SIAM J. CONTROL
Vol. 10, No. 1, February 1972

STABILITY CONDITIONS DERIVED FROM SPECTRAL THEORY: DISCRETE SYSTEMS WITH PERIODIC FEEDBACK*

JON H. DAVIS†

Abstract. By use of the spectral theory of linear operators, necessary and sufficient conditions are derived for the stability of a class of discrete time feedback systems with a periodically time-varying feedback gain. These conditions involve a Nyquist plot for an equivalent time-invariant system which may be determined from the frequency response function of the system under consideration. In the special case of a scalar-input scalar-output system, these conditions may be given in a particularly simple form.

Introduction. In this paper we consider the stability properties of feedback loops formed by the interconnection of two linear elements (see Fig. 1).

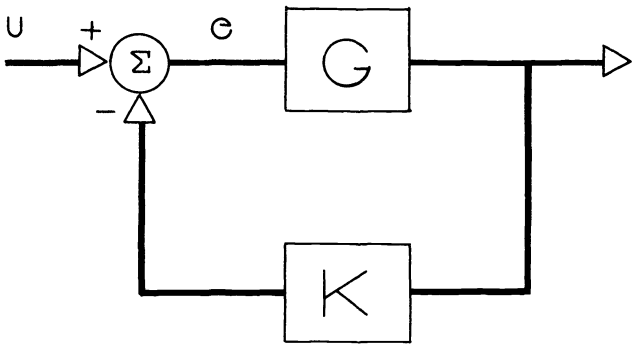


FIG. 1. *Basic feedback loop*

Stability is defined in the input-output sense following the work of Zames [6], and under the assumption that **G** and **K** represent bounded linear operators on a Banach space defined on a half-axis in time, we show in § 1 that the stability (and instability) properties of our feedback system are determined completely by the location of the *spectrum* of the bounded linear operator **KG** acting in a Banach space defined on a half-axis in time. This result we illustrate by giving a generalization of the Nyquist criterion to vector-input vector-output systems.

Using the framework of § 1, we derive a new and explicit criterion which provides a necessary and sufficient condition for the stability of a linear discrete system with a periodic feedback gain. The class of systems to which these results are applicable is limited to those in which the period of the feedback element is an integral multiple of sampler period.

1. Summary of basic results. In this section upper-case script letters ($\mathscr{X}$, $\mathscr{Y}$, etc.) are used to denote Banach spaces defined on a half-axis in time, and lower-case italic letters (x, y, e, etc.) are used to denote elements of such spaces. Operators

* Received by the editors December 16, 1969, and in final revised form February 9, 1971.

† Department of Mathematics, Queen's University, Kingston, Ontario, Canada. This work was supported in part by the National Aeronautics and Space Administration under Grant NGL-22-009-124 with the Electronic Systems Laboratory of the Massachusetts Institute of Technology.

are denoted by boldface upper-case letters. Finite-dimensional vectors and matrices are indicated by a single underline and double underline respectively.

We consider the stability of a certain class of linear feedback systems described by the equations

$$y = \mathbf{G}e, \qquad e = u - \mathbf{K}y. \tag{1}$$

Following Zames [6], we make the following definition of stability.

DEFINITION. The feedback system defined by the functional equation $\{(\mathbf{I} + \mathbf{KG})e = u\}$ is said to be *bounded-input bounded-output stable in the $\mathscr{X}$-sense*, if there exists a constant M such that

$$\|e\|_{\mathscr{X}} \leqq M\|u\|_{\mathscr{X}}$$

for all possible inputs $u \in \mathscr{X}$. If no such M exists, the system is called unstable.

We consider only systems for which the finite time truncated loop equations $\{(\mathbf{I} + \mathbf{KG})e = u,\ 0 \leqq t \leqq T\}$ have a unique solution for each truncation time T. Details on this restriction may be found, for example, in [13], together with a proof of the following theorem. This theorem follows readily from the definition and properties of the *spectrum* of a linear operator. (See, for example, [2].)

THEOREM 1. *Let* $\mathbf{G}$ *and* $\mathbf{K}$ *be bounded linear operators from* $\mathscr{X}$ *to* $\mathscr{Y}$ *and* $\mathscr{Y}$ *to* $\mathscr{X}$ *respectively. Then the feedback system described by the functional equation* $\{(\mathbf{I} + \mathbf{KG})e = u\}$ *is bounded-input bounded-output stable in the* $\mathscr{X}$*-sense if and only if the point* -1 *does not belong to the spectrum of the bounded linear operator* $\mathbf{KG}$ *acting on* $\mathscr{X}$.

Theorem 1 may be applied to derive a generalization of the Nyquist criterion to vector-input vector-output systems. This is accomplished by employing results obtained by Gohberg, Krein, and others [3], [4], [5] on the spectrum of matrix convolution operators. Details of the proofs are given in [13], and the results are stated below since they are needed for § 2.

Notation. We let E^+ denote any one of the following half-axis Banach spaces:

1. $L^p(0, \infty)$, $1 \leqq p \leqq \infty$,
2. $M_c^+(0, \infty)$, the continuous $L^\infty(0, \infty)$-functions,
3. $M_u^+(0, \infty)$, the uniformly continuous $L^\infty(0, \infty)$-functions,
4. $C^+(0, \infty)$, the continuous functions for which the limit as $t \to \infty$ exists, or
5. $C_0^+(0, \infty)$, the subset of $C^+(0, \infty)$ for which the limit as $t \to \infty$ is zero.

By $E_{(n)}^+$ is meant the space of n-vector functions, each element of which belongs to E^+.

THEOREM 2. *Let* $\underline{\underline{G}}(\cdot)$ *be an* $n \times m$ *matrix function, each of whose elements belongs to* $L^1(0, \infty)$, *and let* $\underline{\underline{K}}$ *be a constant* $m \times n$ *matrix. Then the feedback system defined by the equations*

$$\begin{aligned} \underline{y} &= \int_0^t \underline{\underline{G}}(t - s)\underline{e}(s)\, ds, \\ \underline{e} &= \underline{u} - \underline{\underline{K}}\underline{y}, \qquad t \geqq 0, \end{aligned} \tag{2}$$

is bounded-input bounded-output stable on each of the spaces $E_{(m)}^+$ *if and only if the following conditions are satisfied*:

$$\text{(i)} \qquad \det(\underline{\underline{I}} + \underline{\underline{K}}\hat{\underline{\underline{G}}}(i\omega)) \neq 0, \qquad -\infty \leqq \omega \leqq \infty,$$

(ii) $$\kappa = \frac{1}{2\pi}\int_{-\infty}^{\infty} d_\omega \arg\left(\det\left(\underline{\underline{I}} + \underline{\underline{K}}\hat{\underline{\underline{G}}}(i\omega)\right)\right) = 0.$$

In the above, $\hat{\underline{\underline{G}}}(\cdot)$ is the Laplace transform of the matrix function $\underline{\underline{G}}(\cdot)$.

Remarks. Theorem 2 is more easily stated as the condition that the Nyquist locus of the function $\det(\underline{\underline{I}} + \underline{\underline{K}}\hat{\underline{\underline{G}}}(s))$ neither intersect nor encircle the origin of the complex plane. Since the Laplace transform of an $L^1(0, \infty)$-function is analytic in the right half-plane, this is equivalent to the condition that $\det(\underline{\underline{I}} + \underline{\underline{K}}\hat{\underline{\underline{G}}}(s))$ have no zero in the region $\mathrm{Re}(s) \geqq 0$. The sufficiency of the latter condition was previously shown by Desoer and Wu [12].

Theorem 2 has been given in a form directly applicable to continuous time feedback systems. However, as is discussed in Gohberg and Krein [4], completely analogous results hold for discrete time systems. As would be expected, z-transforms are substituted for Laplace transforms, and the half-plane $\mathrm{Re}(s) \geqq 0$ is replaced by the region $|z| \geqq 1$. Since we shall need the discrete version of Theorem 2 for the specific problem considered in the following section, we state it below for completeness.

THEOREM 2′. *Let $\{\underline{\underline{G}}_i\}_{i=1}^{\infty}$ be an $n \times m$ matrix sequence such that $\sum_i |G_i^{kj}| < \infty$, $1 \leqq k \leqq n$, $1 \leqq j \leqq m$. Let $\underline{\underline{K}}$ be a constant $m \times n$ matrix. Then the feedback system defined by the equations*

(3) $$\underline{y}_k = \sum_{j=0}^{k-1} \underline{\underline{G}}_{k-j}\underline{e}_j,$$
$$\underline{e}_k = \underline{u}_k - \underline{\underline{K}}\underline{y}_k, \qquad k \geqq 0,$$

is bounded-input bounded-output stable in the sense of $l^+_{p(m)}$, $p \geqq 1$, $C^+_{(m)}$, and $C^+_{0(m)}$ if and only if the following conditions hold:

(i) $$\det(\underline{\underline{I}} + \underline{\underline{K}}\hat{\underline{\underline{G}}}(e^{i\theta})) \neq 0, \qquad 0 \leqq \theta \leqq 2\pi,$$

(ii) $$\kappa = \frac{1}{2\pi}\int_0^{2\pi} d_\theta \arg\left(\det\left(\underline{\underline{I}} + \underline{\underline{K}}\hat{\underline{\underline{G}}}(e^{i\theta})\right)\right) = 0.$$

Here $\hat{\underline{\underline{G}}}(z) = \sum_{i=1}^{\infty} \underline{\underline{G}}_i z^{-i}$ is the z-transform matrix of the sequence $\{\underline{\underline{G}}_i\}_{i=1}^{\infty}$, which we shall sometimes call the frequency response function.

2. Stability of discrete systems with periodic feedback. As an example of the application of the mathematical results obtained above we consider a vector-input vector-output linear discrete system with periodic feedback gain (Fig. 2).

The equations of the system are as follows:

(4) $$\underline{y}_k = \sum_{j=0}^{k-1} \underline{\underline{G}}_{k-j}\underline{e}_j,$$
$$\underline{e}_k = -\underline{\underline{F}}(k)\underline{y}_k + \underline{u}_k, \qquad k = 0, 1, 2, \cdots.$$

We assume that $\sum_{i=1}^{\infty} |G_i^{jk}| < \infty$, i.e., each element of the $\underline{\underline{G}}$ matrix sequence belongs to l_1^+, and that the feedback gain satisfies $\underline{\underline{F}}(k+n) = \underline{\underline{F}}(k)$ for all k. The sequences $\underline{u}$ and $\underline{e}$ are m-vectors, $\underline{y}$ is an r-vector so that the $\underline{\underline{G}}$ matrices are of dimension $r \times m$, and the $\underline{\underline{F}}$ matrices are of dimension $m \times r$.

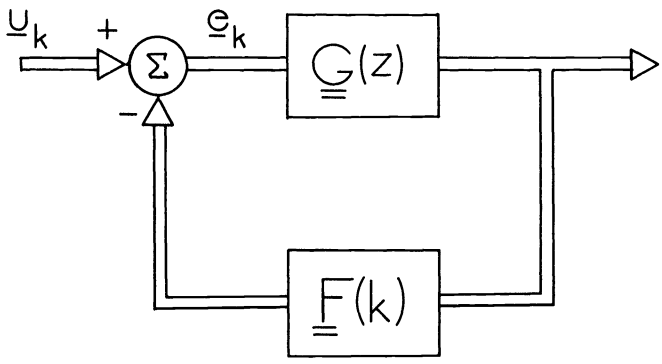


FIG. 2. *Discrete system with periodic feedback*

For reasons of convenience of notation and exposition we choose l_2^+ for the half-axis Banach space, and we derive a necessary and sufficient condition for bounded-input bounded-output stability in the l_2^+-sense. We shall do this by transforming the original system of Fig. 2 into an equivalent system of n-times the dimension of the original, and to which Theorem 2′ of § 1 is applicable.

We consider now the action of the periodic feedback gain operator $\mathbf{F}$ on the space l_2^+. If a sequence $\{\underline{y}_i\}_{i=0}^{\infty}$ has an l.i.m. z-transform

$$\hat{\underline{y}}(z) = \sum_{i=0}^{\infty} \underline{y}_i z^{-i},$$

then the sequence $\{\underline{\underline{F}}(i)\underline{y}_i\}_{i=0}^{\infty}$ has the l.i.m. z-transform

$$(\hat{\underline{\mathbf{F}}}\underline{y})(z) = \sum_{i=0}^{\infty} \underline{\underline{F}}(i)\underline{y}_i z^{-i}.$$

Since the gain matrix $\underline{\underline{F}}(\cdot)$ is periodic of period n, every nth vector in the sequence $\{\underline{y}_i\}_{i=0}^{\infty}$ is multiplied by the same $\underline{\underline{F}}(i)$. This suggests that we split the sequences in the loop into the (orthogonal) direct sum of n copies of l_2^+ by considering the n subsequences whose indices are congruent (mod n) to the integers $0, 1, \cdots, n-1$. This allows us to consider l_2^+ as an orthogonal direct sum of n subspaces each of which is invariant relative to the action of a time-invariant feedback gain operator.

Define the projection operator $\mathbf{P}_i : l_2^+ \to l_2^+$ by the relation

$$\mathbf{P}_i\{\underline{y}_j\}_{j=0}^{\infty} = \{\underline{y}_j^{(i)}\}_{j=0}^{\infty},$$

where

$$\underline{y}_j^{(i)} = \begin{cases} \underline{y}_j & \text{if } i \equiv j \pmod{n}, \\ \underline{0} & \text{if } i \not\equiv j \pmod{n}. \end{cases}$$

Clearly $\mathbf{P}_i\mathbf{P}_j = \delta_{ij}\mathbf{P}_i$, and the $\{\mathbf{P}_i\}_{i=0}^{n-1}$ form a resolution of the identity

$$\mathbf{I} = \mathbf{P}_0 + \mathbf{P}_1 + \cdots + \mathbf{P}_{n-1}. \tag{5}$$

We write the system equations (4) in the form

$$(\mathbf{I} + \mathbf{FG})e = u, \tag{4a}$$

and using (5), we have

$$[\mathbf{I} + \mathbf{F}(\mathbf{P}_0 + \mathbf{P}_1 + \cdots + \mathbf{P}_{n-1})\mathbf{G}(\mathbf{P}_0 + \mathbf{P}_1 + \cdots + \mathbf{P}_{n-1})]e = u.$$

Now we note that $\mathbf{FP}_i = \mathbf{F}(i)\mathbf{P}_i = \mathbf{P}_i\mathbf{F}(i)\mathbf{P}_i$, where $\mathbf{F}(i)$ is the operator on l_2^+ representing multiplication by the ith value of the feedback gain matrix. Therefore,

$$[\mathbf{I} + (\mathbf{F}(0)\mathbf{P}_0 + \mathbf{F}(1)\mathbf{P}_1 + \cdots + \mathbf{F}(n-1)\mathbf{P}_{n-1})\mathbf{G}(\mathbf{P}_0 + \cdots + \mathbf{P}_{n-1})]e = u.$$

Applying successively the projection operators $\mathbf{P}_0, \cdots, \mathbf{P}_{n-1}$ to the above equation, we get

$$\mathbf{P}_j e + \mathbf{P}_j\mathbf{F}(j)\mathbf{P}_j[\mathbf{P}_j\mathbf{G}\mathbf{P}_0\mathbf{P}_0 e + \mathbf{P}_j\mathbf{G}\mathbf{P}_1\mathbf{P}_1 e + \cdots + \mathbf{P}_j\mathbf{G}\mathbf{P}_{n-1}\mathbf{P}_{n-1}e] = \mathbf{P}_j u,$$

where $j = 0, 1, \cdots, n-1$. This may be rewritten as

$$\left\{\mathbf{I} + \begin{bmatrix} \mathbf{F}(0) & & & \\ & \mathbf{F}(1) & & \\ & & \ddots & \\ & & & \mathbf{F}(n-1) \end{bmatrix} \begin{bmatrix} \mathbf{P}_0\mathbf{G}\mathbf{P}_0 & \cdots & \mathbf{P}_0\mathbf{G}\mathbf{P}_{n-1} \\ \vdots & & \vdots \\ \mathbf{P}_{n-1}\mathbf{G}\mathbf{P}_0 & \cdots & \mathbf{P}_{n-1}\mathbf{G}\mathbf{P}_{n-1} \end{bmatrix}\right\} \cdot \begin{bmatrix} \mathbf{P}_0 e \\ \vdots \\ \mathbf{P}_{n-1}e \end{bmatrix} = \begin{bmatrix} \mathbf{P}_0 u \\ \vdots \\ \mathbf{P}_{n-1}u \end{bmatrix}. \tag{6}$$

Since the $\{\mathbf{P}_i\}$ are a set of orthogonal projections satisfying (5), the system (6) is completely equivalent to the original system.

It can be ascertained that the matrix operator $\{\mathbf{P}_i\mathbf{G}\mathbf{P}_j\}$ is equivalent to a convolution operator acting from $l_{2(m)}^+$ to $l_{2(r)}^+$, where these spaces are simply obtained by reindexing the projected original space.[1]

[1] To see this we consider a typical operator $\mathbf{P}_m\mathbf{G}\mathbf{P}_l$. If a sequence $\{\underline{e}_i\}_{i=0}^{\infty}$ has an l.i.m. z-transform

$$\hat{\underline{e}}(z) = \sum_{i=0}^{\infty} \underline{e}_i z^{-i},$$

then the l.i.m. transform of $\mathbf{P}_l\underline{e}$ is

$$\widehat{\mathbf{P}_l\underline{e}}(z) = \sum_{j=0}^{\infty} \underline{e}_{jn+l} z^{-(jn+l)}.$$

The sequence $\{(\mathbf{G}\mathbf{P}_l\underline{e})_i\}_{i=0}^{\infty}$ has l.i.m. z-transform

$$\widehat{\mathbf{G}\mathbf{P}_l\underline{e}}(z) = \sum_{i,j=0}^{\infty} \underline{G}_i \underline{e}_{jn+l} z^{-(jn+l+i)}. \tag{7}$$

If $\underline{y} = \mathbf{P}_m\mathbf{G}\mathbf{P}_l\underline{e}$, then since $\underline{y}$ belongs to the mth subspace,

$$\hat{\underline{y}}(z) = \sum_{q=0}^{\infty} \underline{y}_{qn+m} z^{-(qn+m)}.$$

Identifying the coefficients of $z^{-(qn+m)}$ in (7), we find that the only contribution is from terms where

$$(jn + l + i) = qn + m,$$

or

$$i \equiv (m - l) \bmod n.$$

If we let $y^{(j)}$ denote the reindexed sequence $\mathbf{P}_j y$, and let $\mathbf{G}_{ml}$ denote the convolution operator $\mathbf{P}_m \mathbf{G} \mathbf{P}_l$, then the original system may be represented in the equivalent form

$$\left\{ \mathbf{I} + \begin{bmatrix} \mathbf{F}(0) & & & \\ & \mathbf{F}(1) & & \\ & & \ddots & \\ & & & \mathbf{F}(n-1) \end{bmatrix} \begin{bmatrix} \mathbf{G}_{00} & \cdots & \mathbf{G}_{0(n-1)} \\ \vdots & & \vdots \\ \mathbf{G}_{(n-1)0} & \cdots & \mathbf{G}_{(n-1)(n-1)} \end{bmatrix} \right\} \cdot \begin{bmatrix} e^{(0)} \\ e^{(1)} \\ \vdots \\ e^{(n-1)} \end{bmatrix} = \begin{bmatrix} u^{(0)} \\ \vdots \\ u^{(n-1)} \end{bmatrix} \tag{9}$$

or

$$[\mathbf{I} + \mathbf{K}_{\mathbf{F}} \mathbf{\Gamma}] \underline{e} = \underline{u}.$$

The system (9) has an interpretation as a feedback system with an $rn \times mn$ convolution operator in the forward loop, and an $mn \times rn$ constant matrix in the feedback loop.

The z-transform of the sequence corresponding to the convolution operator $\mathbf{\Gamma}$ in (9) may be found from the z-transform of the original system (4). If the original z-transform matrix is written in the form

$$\hat{\underline{G}}(z) = \hat{\underline{G}}_0(z^n) + z^{-1}\hat{\underline{G}}_1(z^n) + \cdots + z^{-(n-1)}\hat{\underline{G}}_{n-1}(z^n),$$

then the z-transform matrix which corresponds to the convolution $\mathbf{\Gamma}$ may be written as

$$\hat{\mathscr{G}}(z) = \begin{bmatrix} \hat{\underline{G}}_0(z) & \frac{1}{z}\hat{\underline{G}}_{n-1}(z) & \cdots & \frac{1}{z}\hat{\underline{G}}_1(z) \\ \hat{\underline{G}}_1(z) & & & \vdots \\ \vdots & & & \frac{1}{z}\hat{\underline{G}}_{n-1}(z) \\ \hat{\underline{G}}_{n-1}(z) & & \cdots & \hat{\underline{G}}_0(z) \end{bmatrix}. \tag{10}$$

Footnote [1] *continued*

Therefore,

$$\widehat{\mathbf{P}_m \mathbf{G} \mathbf{P}_l \underline{e}}(z) = \sum_{\substack{i \equiv (m-l) \bmod n \\ j}} \underline{G}_i \underline{e}_{jn+l} z^{-(jn+l+i)} = \sum_{q,j} \underline{G}_{qn+(m-l)} \underline{e}_{jn+l} z^{-(jn+qn+m)}. \tag{8}$$

A sequence $\mathbf{P}_l \underline{e}$ has the form

$$\{0, 0, \cdots, e_l, 0, \cdots, 0, e_{l+n}, 0, \cdots, 0, e_{l+2n}, 0, \cdots\}.$$

If we reindex the sequences $\mathbf{P}_j \underline{e}$, $j = 0, \cdots, n-1$, counting only (potentially) nonzero terms, then we see from (8) that the operator $\mathbf{P}_m \mathbf{G} \mathbf{P}_l : \mathbf{P}_l l_2^+ \to \mathbf{P}_m l_2^+$ is simply a convolution relative to the new indices. The sequence which is the kernel of the convolution is obtained from the original sequence by taking every nth member of the original, starting with the $(m-l)$th.

This representation follows readily from the observation made above that the convolution operator $\mathbf{P}_m\mathbf{G}\mathbf{P}_l$ corresponds to a sequence obtained from the original by taking every nth member of the original sequence starting with the $(m-l)$th.

In the practically important case where the original system (4) has a representation of the form

$$\begin{aligned} \underline{x}_{k+1} &= \underline{\underline{A}}\underline{x}_k + \underline{\underline{B}}\underline{e}_k, \\ \underline{y}_k &= \underline{\underline{C}}\underline{x}_k, \\ \underline{e}_k &= -\underline{\underline{F}}(k)\underline{y}_k + \underline{u}_k, \end{aligned} \tag{4'}$$

then $\hat{\underline{\underline{G}}}(z)$ has an explicit representation of the form

$$\hat{\underline{\underline{G}}}(z) = \begin{bmatrix} \underline{\underline{C}}(\underline{\underline{I}}z - \underline{\underline{A}}^n)^{-1}\underline{\underline{A}}^{n-1}\underline{\underline{B}} & \underline{\underline{C}}(\underline{\underline{I}}z - \underline{\underline{A}}^n)^{-1}\underline{\underline{A}}^{n-2}\underline{\underline{B}} & \cdots & \underline{\underline{C}}(\underline{\underline{I}}z - \underline{\underline{A}}^n)^{-1}\underline{\underline{B}} \\ z\underline{\underline{C}}(\underline{\underline{I}}z - \underline{\underline{A}}^n)^{-1}\underline{\underline{B}} & & & \\ z\underline{\underline{C}}(\underline{\underline{I}}z - \underline{\underline{A}}^n)^{-1}\underline{\underline{A}}\underline{\underline{B}} & & & \\ \vdots & & & \\ z\underline{\underline{C}}(\underline{\underline{I}}z - \underline{\underline{A}}^n)^{-1}\underline{\underline{A}}^{n-2}\underline{\underline{B}} & \cdots & z\underline{\underline{C}}(\underline{\underline{I}}z - \underline{\underline{A}}^n)^{-1}\underline{\underline{B}} & \underline{\underline{C}}(\underline{\underline{I}}z - \underline{\underline{A}}^n)^{-1}\underline{\underline{A}}^{n-1}\underline{\underline{B}} \end{bmatrix}. \tag{11}$$

Notation. If $\underline{\underline{F}}(\cdot)$ is an $m \times r$ matrix function defined on the integers and of period n, then we let

$$\underline{\underline{K}}_F = \begin{bmatrix} \underline{\underline{F}}(0) & 0 & 0 & \cdots & \\ \vdots & \underline{\underline{F}}(1) & 0 & & \\ 0 & & \ddots & & \vdots \\ \vdots & & & \ddots & \\ \vdots & & \cdots & & \underline{\underline{F}}(n-1) \end{bmatrix}$$

denote the block diagonal $mn \times rn$ matrix constructed from the $\underline{\underline{F}}(i)$.

THEOREM 3. *Let $\hat{\underline{\underline{G}}}(z)$ be the z-transform of a sequence of matrices $\{\underline{\underline{G}}\}_{i=1}^{\infty}$ of dimension $r \times m$, such that $\sum_{i=1}^{\infty} |G_i^{kj}| < \infty$ for $1 \leqq k \leqq r$, $1 \leqq j \leqq m$. Let $\underline{\underline{F}}(\cdot)$ be a periodic matrix function of dimension $m \times r$ and period n. Then the feedback system described by the equations*

$$\begin{aligned} \underline{y}_k &= \sum_{j=0}^{k-1} \underline{\underline{G}}_{k-j}\underline{e}_j, \\ \underline{e}_k &= \underline{u}_k - \underline{\underline{F}}(k)\underline{y}_k, \qquad k \geqq 0, \end{aligned} \tag{4}$$

is bounded-input bounded-output stable in the sense of $l_{p(m)}^+$, $p \geqq 1$, $C_{(m)}^+$, and $C_{0(m)}^+$ if and only if the following conditions are satisfied:

(i) $$\det(\underline{\underline{I}} + \underline{\underline{K}}_F\hat{\underline{\underline{G}}}(e^{i\theta})) \neq 0, \qquad 0 \leqq \theta \leqq 2\pi,$$

(ii) $$\kappa = \frac{1}{2\pi}\int_0^{2\pi} d_\theta \arg(\det(\underline{\underline{I}} + \underline{\underline{K}}_F\hat{\underline{\underline{G}}}(e^{i\theta})) = 0.$$

Here, if $\hat{\underline{\underline{G}}}(z) = \hat{\underline{\underline{G}}}_0(z^n) + z^{-1}\hat{\underline{\underline{G}}}_1(z^n) + \cdots + z^{-(n-1)}\hat{\underline{\underline{G}}}_{n-1}(z^n)$,

$$\hat{\underline{\underline{\mathcal{G}}}}(z) \triangleq \begin{bmatrix} \hat{\underline{\underline{G}}}_0(z) & \frac{1}{z}\hat{\underline{\underline{G}}}_{n-1}(z) & \cdots & \frac{1}{z}\hat{\underline{\underline{G}}}_1(z) \\ \hat{\underline{\underline{G}}}_1(z) & \hat{\underline{\underline{G}}}_0(z) & & \\ \vdots & \ddots & \ddots & \\ & & & \frac{1}{z}\hat{\underline{\underline{G}}}_{n-1}(z) \\ \hat{\underline{\underline{G}}}_{n-1}(z) & \cdots & \hat{\underline{\underline{G}}}_1(z) & \hat{\underline{\underline{G}}}_0(z) \end{bmatrix},$$

and $\underline{\underline{K}}_F$ is as defined above.

Proof. By the construction of $\underline{\underline{K}}_F$ and $\hat{\underline{\underline{\mathcal{G}}}}(z)$, the original system (4) and a system consisting of a convolution operator $\boldsymbol{\Gamma}$ having a z-transform matrix equal to $\hat{\underline{\underline{\mathcal{G}}}}(z)$ and a (constant) feedback gain $\underline{\underline{K}}_F$ are completely equivalent. By Theorem 2′, the stability of the second system is determined by the above conditions, and hence the stability of the original system is determined as well.

Remarks. As was mentioned in § 1 above, the conditions for stability (and instability) may be stated in terms of the Nyquist locus of a certain function of the complex variable z, in this case the function $\det(\underline{\underline{I}} + \underline{\underline{K}}_F\hat{\underline{\underline{\mathcal{G}}}}(z))$. In general, the determinant involved is such a complicated expression that it is difficult to evaluate and interpret. However, in the case of a scalar-input scalar-output system (where $\underline{\underline{K}}_F$ is diagonal, and $\hat{\underline{\underline{\mathcal{G}}}}(z)$ square) it is possible to make some progress.

An alternative expression of the condition of Theorem 3 may be derived from the following result, due to W. E. Roth [8].

THEOREM. *If $\underline{\underline{A}}$ is a $p \times p$ matrix with elements in the field F, and $\underline{\underline{D}}_i = -\underline{b}\underline{c}_i'$, and if the characteristic polynomial of $\underline{\underline{A}}_i = \underline{\underline{A}} + \underline{\underline{D}}_i$ is*

$$\det(\underline{\underline{I}}x - \underline{\underline{A}}_i) = m_{i,0} + m_{i,1}x + m_{i,2}x^2 + \cdots + m_{i,(q-1)}x^{q-1},$$

where $m_{i,(j-1)}$, $i, j = 1, 2, \cdots, q$, are polynomials in x^q with coefficients in F, then the characteristic polynomial of the product $\underline{\underline{\Phi}} = \underline{\underline{A}}_1\underline{\underline{A}}_2 \cdots \underline{\underline{A}}_q$ is given by $(-1)^{(q-1)p}\Delta(x)$, where

$$\Delta(x^q) = \det \begin{bmatrix} m_{1,0} & m_{1,q-1}x^{q-1} & m_{1,q-2}x^{q-2} & \cdots & m_{1,1}x \\ m_{2,1}x & m_{2,0} & m_{2,q-1}x^{q-1} & \cdots & m_{2,2}x^2 \\ \vdots & \vdots & \vdots & & \vdots \\ m_{q,q-1}x^{q-1} & m_{q,q-2}x^{q-2} & \cdots & & m_{q,0} \end{bmatrix}.$$

Proof. See [8].

The above theorem may be used to calculate the characteristic polynomial of the transition matrix $\underline{\underline{\Phi}}(n, 0)$ for a linear discrete system described by the equation

$$\underline{x}_{k+1} = [\underline{\underline{A}} - \underline{b}\underline{c}'(k)]\underline{x}_k + \underline{b}u_k, \tag{12}$$

where the vector of functions $\underline{c}'(k)$ satisfies $\underline{c}'(k + n) = \underline{c}'(k)$.

As is well known, if $\underline{c}'(k)(\underline{\underline{I}}z - \underline{\underline{A}})^{-1}\underline{b} = q_k(z)/p(z)$, then

$$\det(\underline{\underline{I}}z - \underline{\underline{A}} + \underline{b}\underline{c}'(k)) = p(z) + q_k(z).$$

This means that the polynomials in Roth's theorem above may be readily found. In the case where $\underline{c}'(k) = f(k)\underline{c}'$, and $\hat{g}(z) = \underline{c}'(\underline{I}z - \underline{\underline{A}})^{-1}\underline{b} = q(z)/p(z)$, the polynomials $m_{i,(j-1)}$ may be found simply from $q(z)$ and $p(z)$. Let $q_j(z)$ be the polynomial made from $q(z)$ by retaining only those terms whose powers are congruent to $j \pmod n$, where n is the period of the system. Then we have

$$z^{(j-1)}m_{i,(j-1)}(z) = p_{j-1}(z) + f(i)q_{j-1}(z).$$

In this case the determinant of the matrix of polynomials occurring in Roth's theorem may be written as

$$\Delta(z^n) = \det \left\{ \begin{bmatrix} p_0(z) & p_{n-1}(z) & p_{n-2}(z) & \cdots & p_1(z) \\ p_1(z) & p_0(z) & & & \\ p_2(z) & p_1(z) & & & \\ \vdots & & \ddots & & \\ p_{n-1}(z) & & & & p_0(z) \end{bmatrix} + \begin{bmatrix} f_{(0)} & & & \\ & & & \\ & & & \\ & & & f_{(n-1)} \end{bmatrix} \begin{bmatrix} q_0(z) & q_{n-1}(z) & \cdots & q_1(z) \\ q_1(z) & & & \\ \vdots & & & \\ q_{n-1}(z) & & \cdots & q_0(z) \end{bmatrix} \right\}. \tag{13}$$

Since the difference equation

$$\underline{x}_{k+1} = [\underline{\underline{A}} - f(k)\underline{b}\underline{c}']\underline{x}_k \tag{12'}$$

will be asymptotically stable if and only if all of the eigenvalues of its transition matrix $\underline{\underline{\Phi}}(n, 0)$ lie inside the unit disc, the condition for asymptotic stability may be expressed by requiring that the polynomial $\Delta(x^n)$ have no zero in the region $|x| \geqq 1$. Hence we have the following.

THEOREM 3'. *The difference equation*

$$\underline{x}_{k+1} = [\underline{\underline{A}} - f(k)\underline{b}\underline{c}']\underline{x}_k, \qquad k \geqq 0,$$

with $f(k + n) = f(k)$, *is asymptotically stable if and only if* $\Delta(x^n)$ *has no zero in the region* $|x| \geqq 1$. *Here* $\Delta(x^n)$ *is as defined in* (13) *above.*

Remarks. Theorem 3' has the advantage that the function $\Delta(x^n)$ may be determined readily from the frequency response function (provided it is a rational function) of the original system. There is no need to compute the expanded matrix $\underline{\underline{\hat{G}}}(z)$ of Theorem 3. Theorem 3' is also interesting in that it involves the polynomials which determine the "time-varying root locus" (i.e., the polynomials $p(z) + f(k)q(z)$, $k = 0, \cdots, n - 1$) of the closed loop of Fig. 3.

The relationship between Theorem 3 and Theorem 3' is not immediately evident, although it is relatively easy to establish by the use of the following lemma.

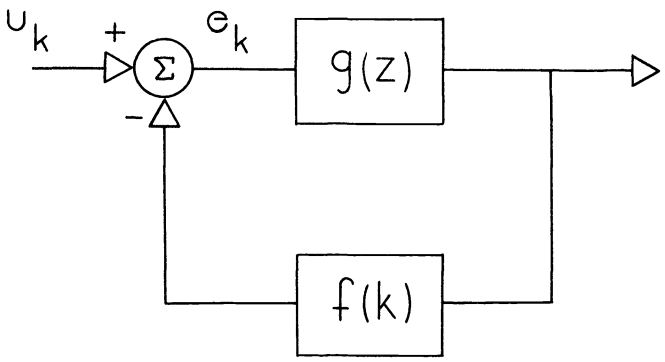


FIG. 3. *Scalar-input scalar-output system*

LEMMA. *The matrix* $\hat{\underline{\underline{G}}}(z)$ *corresponding to a scalar frequency response function* $\hat{g}(z)$ *has a complete set of eigenvectors*

$$\underline{v}_j = \begin{bmatrix} 1 \\ z^{1/n}\omega_j \\ z^{2/n}\omega_j^2 \\ \vdots \\ z^{(n-1)/n}\omega_j^{n-1} \end{bmatrix}, \qquad j = 0, 1, \cdots, n-1,$$

where $z^{1/n}$ *denotes the principal value of the n-th root of the complex variable* z, *and* $\omega_j = \exp(2\pi ij/n)$ *denotes an n-th root of unity. Furthermore, the eigenvalue corresponding to* $\underline{v}_j$ *is* $\lambda_j = \hat{g}(\omega_j z^{1/n})$.

Proof. If $\hat{g}(z) = \hat{g}_0(z^n) + z^{-1}\hat{g}_1(z^n) + \cdots + z^{-(n-1)}\hat{g}_{n-1}(z^n)$ is the frequency response function, then

$$(10') \qquad \hat{\underline{\underline{G}}}(z) = \begin{bmatrix} \hat{g}_0(z) & \frac{1}{z}\hat{g}_{n-1}(z) & \cdots & \frac{1}{z}\hat{g}_1(z) \\ \hat{g}_1(z) & \hat{g}_0(z) & & \\ \vdots & & \ddots & \\ \hat{g}_{n-1}(z) & & & \hat{g}_0(z) \end{bmatrix}.$$

Evaluating $\hat{g}(\cdot)$ at the point $z^{1/n}\omega_j$ shows that

$$\hat{g}(\omega_j z^{1/n}) = \hat{g}_0(z) + (\omega_j z^{1/n})^{-1}\hat{g}_1(z) + \cdots + (\omega_j z^{1/n})^{-(n-1)}\hat{g}_{n-1}(z),$$

and this expression is the key to a straightforward verification.

COROLLARY. $\hat{\underline{\underline{G}}}(z)$ *may be expressed as*

$$\hat{\underline{\underline{G}}}(z) = \operatorname{diag}(1, z^{1/n}, \cdots, z^{(n-1)/n}) \cdot \underline{\underline{\Omega}} \cdot \operatorname{diag}(\hat{g}(\omega_0 z^{1/n}) \cdots \cdot \hat{g}(\omega_{n-1} z^{1/n})) \cdot \underline{\underline{\Omega}}^* \cdot \operatorname{diag}(1, z^{-1/n}, \cdots, z^{-(n-1)/n}),$$

where

$$\underline{\underline{\Omega}} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \omega_0 & \omega_1 & & \omega_{(n-1)} \\ \omega_0^2 & \omega_1^2 & & \omega_{(n-1)}^2 \\ \vdots & \vdots & & \vdots \\ \omega_0^{(n-1)} & \omega_1^{(n-1)} & \cdots & \omega_{(n-1)}^{(n-1)} \end{bmatrix} \cdot \frac{1}{\sqrt{n}}$$

is a (unitary) normalized Vandermonde matrix.

Recalling that the spectrum of the convolution operator **G** corresponding to the frequency response function $\hat{g}(z)$ consists of the range of $\hat{g}(z)$ for $|z| \geqq 1$ (see [1]), we may interpret the above lemma in terms of a partition of the original spectrum into n-separate pieces. The result of the lemma should not be too surprising, since Theorem 3 must reduce to just the Nyquist criterion in the case of a constant feedback gain.

The expression $\det(\underline{I} + \underline{K}_F\hat{\underline{\underline{G}}}(z))$ may be manipulated into various equivalent forms by the use of a similarity transformation and the identity

$$\det(\underline{I} + \underline{A}\underline{B}) = \det(\underline{I} + \underline{B}\underline{A}). \tag{14}$$

One such equivalent form is given by the identity

$$\det(\underline{I} + \underline{K}_F\hat{\underline{\underline{G}}}(z)) = \det\{\underline{I} + \underline{\Omega}^*\underline{K}_F\underline{\Omega} \cdot \operatorname{diag}(\hat{g}(\omega_0 z^{1/n}) \cdots \hat{g}(\omega_{n-1} z^{1/n}))\}, \tag{15}$$

which follows from the lemma and (14).

The matrix $\underline{\Omega}$ is the matrix which corresponds to the standard discrete Fourier transform. A short computation shows that the matrix $\underline{\Omega}^*\underline{K}_F\underline{\Omega}$ is simply a Toeplitz matrix made up of the discrete Fourier coefficients of the periodic feedback gain. Special cases of this expression may be derived by employing discrete Fourier series expansions in connection with difference equations of the form

$$\underline{x}_{k+1} = [\underline{A} - f(k)\underline{b}\underline{c}']\underline{x}_k, \qquad f(k+n) = f(k),$$

under the assumption that a Floquet-type representation of the transition matrix exists.

The connection between Theorem 3 and Theorem 3′ may be established by noting that the two matrices of polynomials which occur in the expression (13) are examples of so-called *circulant matrices* [15]. A circulant matrix is a Toeplitz matrix with the additional property that the elements $t_{ij} = t_{(i-j)}$ are such that $t_l = t_m$ if $l \equiv m \pmod{n}$. If $\underline{C}$ is a circulant matrix with entries $c_{ij} = c_{(i-j)}$, then it is easily verified that

$$\underline{\Omega}^*\underline{C}\underline{\Omega} = \operatorname{diag}((c_0 + c_1\omega_j + \cdots + c_{n-1}\omega_j^{n-1})),$$

where $\underline{\Omega}$ is the unitary matrix defined above. Premultiplying the matrix whose determinant in (13) is to be computed by $\underline{\Omega}^*$, and postmultiplying the matrix by $\underline{\Omega}$, shows that

$$\Delta(z^n) = \det[\operatorname{diag}(p(\omega_j z)) + \underline{\Omega}^*\underline{K}_F\Omega \operatorname{diag}(q(\omega_j z))].$$

Comparison of the above with (15) shows that

$$\Delta(z^n) = \prod_{j=0}^{n-1} p(\omega_j z) \cdot \det (\underline{I} + \underline{K}_F \hat{\underline{\mathcal{G}}}(z^n)). \tag{16}$$

We have defined the matrix function $\hat{\underline{\mathcal{G}}}(z)$ only for systems which correspond to bounded operators, which means in the present case that the polynomial $p(z)$ has no zero in the region $|z| \geqq 1$. Equation (16) shows that in this case the functions $\Delta(z^n)$ and $\det (\underline{I} + \underline{K}_F \hat{\underline{\mathcal{G}}}(z^n))$ both have the same number of zeros in the region $|z| \geqq 1$, so that Theorem 3 and Theorem 3′ are equivalent for this case. In the case that $p(z)$ has a zero in the region $|z| \geqq 1$, Theorem 3 is not directly applicable, so that there is no comparison possible.

3. Some general remarks.

1. The results of § 2 represent one of the few examples of a class of feedback systems for which necessary and sufficient stability conditions may be given in a relatively tractable analytical form.

2. The device of "expanding" the original system to a higher-dimensional equivalent one may be applied to put systems with a time-varying gain which has a dominant periodic component into a form in which the circle criterion or similar results may be applied.

3. The methods of § 2 provide necessary and sufficient conditions for the positivity of an operator composed of a periodic gain and a discrete convolution.

4. The extension of the methods of § 2 to the case of continuous time systems is not a straightforward matter. Somewhat delicate problems of analysis occur in the continuous time case, which are avoided in the present problem essentially because of the "closeness" (in terms of general behavior) of discrete time function spaces to finite-dimensional spaces. This extension is an area of current research.

5. The idea of matrix manipulations of the type used to establish the equivalence of Theorem 3 and Theorem 3′ has not been fully exploited. As an example, permutation matrices might be introduced as a similarity transformation in order to study the effects of permuting the order of occurrence of the values assumed by the feedback gain. The combination of permutation matrices and the discrete Fourier matrix $\underline{\Omega}$ encountered above is connected with the so-called "fast Fourier transform" algorithms. See [16].

6. These results have some potential application to sampled continuous time systems. However if the continuous time problem has a periodic feedback gain, then there is an implicit restriction in the problem formulation that the period of the feedback element be an integral multiple of the sampling period.

Acknowledgment. The author would like to thank Professor R. W. Brockett and Professor J. C. Willems for some helpful suggestions made during the course of the work presented here. The author is also indebted to the reviewers for many helpful comments made on an earlier draft of this paper.

REFERENCES

[1] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operations, Part II*, Interscience, New York, 1967.

[2] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, American Mathematical Society, New York, 1957.

[3] M. G. KREIN, *Integral equations on a half-line with kernel depending on the difference of the arguments*, Amer. Math. Soc. Transl. (2), 22 (1962), pp. 163–288.

[4] I. C. GOHBERG AND M. G. KREIN, *Systems of integral equations on a half-line with kernel depending on the difference of the arguments*, Ibid., (2), 14 (1960), pp. 217–287.

[5] E. I. GOLDENHERSHEL, *The spectrum of a Volterra operator of a half-line and the exponential growth of the solution of a system of integral equations of the Volterra type*, Math. Sbornik, 64(106), (1964), no. 1, pp. 115–139.

[6] G. ZAMES, *On the input-output stability of time-varying non-linear feedback systems. Parts I and II*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238, 465–476.

[7] J. C. WILLEMS, *Stability, instability, invertibility and causality*, this Journal, 7 (1969), pp. 645–671.

[8] W. E. ROTH, *On the characteristic polynomial of the product of several matrices*, Proc. Amer. Math. Soc., 7 (1956), pp. 578–582.

[9] T. W. GAMELIN, *Notes on functional analysis*, Course 18.36, Massachusetts Institute of Technology, Cambridge, 1968.

[10] I. C. GOHBERG AND M. G. KREIN, *Fundamental aspects of defect numbers, root numbers, and indices of linear operators*, Amer. Math. Soc. Transl., 13 (1960), pp. 185–264.

[11] R. W. BROCKETT AND H. B. LEE, *Frequency domain instability conditions for time-varying and non-linear systems*, IEEE Proc., 55 (1967), pp. 604–619.

[12] C. A. DESOER AND C. T. WU, *Stability of multiple-loop feedback linear time-invariant systems*, J. Math. Anal. Appl., 23 (1968), pp. 121–129.

[13] J. H. DAVIS, *Stability of linear feedback systems*, Doctoral thesis, Dept. of Mathematics, Massachusetts Institute of Technology, Cambridge, 1970.

[14] J. R. RINGROSE, *Compact linear operators of the Volterra type*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 44–55.

[15] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

[16] P. J. NICHOLSON, *Algebraic theory of the finite Fourier transform*, Doctoral thesis, Dept. of Operations Research, Stanford University, Stanford, Calif., 1969.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

OPTIMAL CONTROL PROBLEMS WITH A SYSTEM OF INTEGRAL EQUATIONS AND RESTRICTED PHASE COORDINATES*

SHENG-CHAO HUANG†

Abstract. The purpose of this paper is to discuss the variational problems and the optimal control problems characterized by a system of nonlinear Volterra integral equations. The necessary conditions are obtained, for cases both with or without restricted phase coordinates, in the integral form of the maximum principle.

1. Introduction. In recent years, many generalizations [1] of the maximum principle have been obtained. We shall mention some which are related to our present paper. Gamkrelidze [2] formulated a general extremal problem in the theory of differential equations and derived necessary conditions for extremality. He also introduced the concept of quasi-convexity in the same paper. Neustadt [3], [4] introduced the concept of first order convex approximations and proved the abstract multiplier rule in a locally convex linear topological space, and then applied these results to a number of variational problems. The most general case has been done by Halkin and Neustadt [5], where a very general maximum principle for optimizing problems over an arbitrary set was given.

In this paper we shall use the multiplier rules developed in [3] to derive necessary conditions for variational problems characterized by nonlinear Volterra integral equations. In order to include as large a class of optimal control problems as possible, we generalize the concept of quasi-convexity [2]. In § 2, we formally state the problems and present the results. These results include the maximum principle in integral form for Problems 1 and 2. The proof of these results is given in § 4. In § 3, we apply the results of § 2 to Problems 3 and 4, optimal control problems both with or without restricted phase coordinates, obtaining the maximum principle in integral form.

Optimal control problems with systems of linear Volterra integral equations were first considered by Friedman [6]. Halanay, applying the multiplier rule due to Hestenes, has also obtained a maximum principle in pointwise form for a generalized Volterra integral equation (see [7]). An optimal control problem with a nonlinear Volterra integral equation, fixed terminal time and restricted phase coordinates has been considered by Vinokurov [11]. In this paper, we approach the problem considerably differently from all papers mentioned above. In Problem 1 and Problem 3, we allow both terminal time and phase coordinates to be free. In Problem 2, and hence Problem 4, we consider a restricted phase coordinates problem with fixed terminal time.

2. Problem statements and the results. Let I denote the compact interval $[t_1, t_2]$, and let $\mathscr{L}^2$ denote the set of all square integrable functions defined on I. Let C denote the normed linear vector space of all absolutely continuous functions from I into R^n with norm defined by

$$\|x\| = \max_{t \in I} |x(t)|$$

* Received by the editors July 15, 1969, and in final revised form April 23, 1971.

† School of Engineering and Applied Science, University of California, Los Angeles, California. Now at Department of Electrical Engineering, Syracuse University, Syracuse, New York 13210.

for $x \in C$, where $|\cdot|$ denotes the Euclidean norm in R^n. Let $\mathscr{H}$ be a given convex subset of C. A convex hull of a set A in any linear space will be denoted by $[A]$.

For every positive integer $v > 0$, P^v will denote the following subset of R^v;

$$P^v = \left\{ \beta = (\beta^1, \cdots, \beta^v);\ \beta^i \geqq 0,\ \sum_{i=1}^{v} \beta^i = 1 \right\}.$$

The components of a vector $x \in R^n$ will be denoted by superscripts; i.e., the ith component of x will be denoted by x^i. Subscripts will be used to differentiate between the vectors in R^n.

Let G be a given nonempty open set in R^n. Now, let $\mathscr{S}$ denote the linear vector space of all functions $g(x, s, t)$ from $G \times I^2$ into R^n which are of class C^1 with respect to x for each $(s, t) \in I^2$, and, together with $g_x(x, s, t)$, are continuously differentiable with respect to t for each $(x, s) \in G \times I$, and are Borel measureable with respect to $s \in I$ for each $(x, t) \in G \times I$. Let us define a uniformly quasi-convex family of functions as below.

DEFINITION 2.1. A subset $\mathscr{G}$ of $\mathscr{S}$ will be called *uniformly quasi-convex* in I if it satisfies the following conditions:

(a) For each $g \in \mathscr{G}$ and every compact subset X of G, there exists a function $m \in \mathscr{L}^2$ (m may depend on g and X) such that, for all $t \in I$,

$$|g(x, s, t)| \leqq m(s) \tag{2.1}$$

and

$$|g(x, s, t) - g(y, s, t)| \leqq m(s)|x - y| \tag{2.2}$$

for every $x, y \in X$, and all $s \in I$.

(b) For every finite subset $(g_1, \cdots, g_v)$ of $\mathscr{G}$, every compact subset X of G and every real positive number $\eta > 0$, there exist functions $g_\beta \in \mathscr{G}$ defined for every $\beta = (\beta^1, \cdots, \beta^v) \in P^v$ such that the functions

$$\delta^2 g(x, s, t; \beta) = g_\beta(x, s, t) - \sum_{i=1}^{v} \beta^i g_i(x, s, t) \tag{2.3}$$

satisfy the following conditions:

(i)

$$|\delta^2 g(x, s, t; \beta)| \leqq \tilde{m}(s) \tag{2.4}$$

and

$$|\delta^2 g(x, s, t; \beta) - \delta^2 g(y, s, t; \beta)| \leqq \tilde{m}(s)|x - y| \tag{2.5}$$

for every $x, y \in X$, $t \in I$, $s \in I$, and every $\beta \in P^v$, and for some $\tilde{m} \in \mathscr{L}^2$ which may depend upon X and the g_i, but not on η;

(ii)

$$\left| \int_{t_1}^{\tau} \delta^2 g(x, s, t; \beta)\, ds \right| < \eta \tag{2.6}$$

for every $(x, t) \in X \times I$, $\tau \in I$ and every $\beta \in P^v$;

(iii)

$$\sup_{\bar{t}\in I}\left|\int_{t_1}^{\bar{t}}[\delta^2 g(x,s,t;\beta_i)-\delta^2 g(x,s,t;\beta)]\,ds\right|\xrightarrow[\{\beta_i\}\in P^\nu]{\beta_i\to\beta}0 \tag{2.7}$$

for every $(x,t)\in X\times I$, and $\beta\in P^\nu$.

This definition of quasi-convexity is a generalization of a concept which was first introduced by Gamkrelidze [2]. For each fixed $t\in I$, let $\mathscr{G}^t\subset\mathscr{G}$ be a family of functions $g^t(x,s)=g(x,s,t)$. It is obvious that $\mathscr{G}^t$ is quasi-convex in the sense given in [2].

PROBLEM 1. Let I, C, $\mathscr{S}$ be as defined previously, let $\mathscr{H}$ be a preassigned convex subset of C, and let $\mathscr{G}$ be a uniformly quasi-convex subset of $\mathscr{S}$. Let Q be a subset of C such that if $x\in Q$, then for some $h\in\mathscr{H}$ and some $g\in\mathscr{G}$,

$$x(t)=h(t)+\int_{t_1}^{t}g(x(s),s,t)\,ds\quad\text{for all } t\in I. \tag{2.8}$$

Let $B=[t_1,\tau]$ be a convex subset of I such that $t_1<\tau<t_2$.

Let χ_i, $i=-\mu,\cdots,0,\cdots,m$, be a given real-valued, continuously differentiable function defined on $G^2\times I$, and let ϕ_i, $i=-\mu,\cdots,0,\cdots,m$, be functionals on $Q\times I$ such that

$$\phi_i(x,\tau)=\chi_i(x(t_1),x(\tau),\tau),\qquad i=-\mu,\cdots,0,\cdots,m. \tag{2.9}$$

Now, our problem is to find an element $(z,\tau^*)\in Q\times B$ such that $\phi_i(z,\tau^*)\leqq 0$ for $i=-\mu,\cdots,-1$, and $\phi_i(z,\tau^*)=0$ for $i=1,\cdots,m$, and such that

$$\phi_0(z,\tau^*)\leqq\phi_0(x,\tau) \tag{2.10}$$

for all $(x,\tau)\in Q\times B$ which satisfy the relations $\phi_i(x,\tau)\leqq 0$ for $i=-\mu,\cdots,-1$, and $\phi_i(x,\tau)=0$ for $i=1,\cdots,m$.

If (z,τ^*) is a solution to Problem 1, then we shall make the following assumptions:

(a) τ^* is an interior point of B;

(b) $z(t)$ is differentiable at τ^*, and for some $h^*\in\mathscr{H}$ and some $g^*\in\mathscr{G}$, it satisfies the following integral equation:

$$z(t)=h^*(t)+\int_{t_1}^{t}g^*(z(s),s,t)\,ds\quad\text{for all } t,\quad t_1\leqq t\leqq\tau^*; \tag{2.11}$$

(c) the relations

$$\sum_{i=-\mu}^{m}\alpha^i\frac{\partial\chi_i}{\partial\xi_1}=\sum_{i=-\mu}^{m}\alpha^i\frac{\partial\chi_i}{\partial\xi_2}=0,\qquad\sum_{i=-\mu}^{m}\alpha^i\frac{\partial\chi_i}{\partial\tau}=0, \tag{2.12}$$

$$\alpha^i\leqq 0\text{ for } i\leqq 0,\qquad\alpha^i=0\quad\text{for } i<0\quad\text{and such that}\quad\phi_i(z,\tau^*)<0,$$

imply that $\alpha^i=0$, $i=-\mu,\cdots,m$. Here, in (2.12), $\partial\chi_i/\partial\xi_1$ and $\partial\chi_i/\partial\xi_2$ denote the vector with its first n-components and its second n-components, respectively, of the gradient of χ_i evaluated at $(z(t_1),z(\tau^*),\tau^*)$, and $\partial\chi_i/\partial\tau$ is a partial derivative of χ_i with respect to τ also evaluated at $(z(t_1),z(\tau^*),\tau^*)$.

It was shown in [4] that, under the above assumption, (4.48) holds.

We now have the following results.

THEOREM 1. *If (z, τ^*) is a solution to Problem 1, satisfying the previous assumptions, then there exist an n-row vector-valued function ψ defined on (t, τ^*), and real numbers $\alpha^{-\mu}, \cdots, \alpha^0, \cdots, \alpha^m$ such that:*

(a) $$\sum_{i=-\mu}^{m} |\alpha^i| > 0, \qquad \alpha^i \leqq 0 \quad \text{for } i \leqq 0,$$
$$\alpha^i = 0 \quad \text{for } i < 0 \quad \text{and } \phi_i(z, \tau^*) < 0;$$

(b) $$\dot{\psi}(t) = -\psi(t) g_x^*(z(t), t, t) - \int_t^{\tau^*} \psi(\zeta) g_{x\zeta}^*(z(t), t, \zeta)\, d\zeta$$
for almost all $t \in [t_1, \tau^*]$;

(c) $$\psi(t_1) = -\sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1};$$

(d) $$\psi(\tau^*) = \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2};$$

(e) $$\psi(\tau^*)\dot{z}(\tau^*) = -\sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \tau};$$

(f) *$\psi(t)$ is absolutely continuous in $[t_1, \tau^*]$ and is not identical to zero in the interval;*

(g) $$\int_{t_1}^{\tau^*} \psi(s)\dot{h}(s)\, ds \leqq \int_{t_1}^{\tau^*} \psi(s)\dot{h}^*(s)\, ds \qquad \text{for all } h \in \mathscr{H} \text{ such that } h(t_1) = h^*(t_1);$$

(h) $$\int_{t_1}^{\tau^*} \psi(s)\left[g(z(s), s, s) + \int_{t_1}^{s} g_s(z(\zeta), \zeta, s)\, d\zeta\right] ds$$
$$\leqq \int_{t_1}^{\tau^*} \psi(s)\left[g^*(z(s), s, s) + \int_{t_1}^{s} g_s^*(z(\zeta), \zeta, s)\, d\zeta\right] ds \quad \text{for all } g \in \mathscr{G}.$$

Note that condition (b) is the adjoint equation, conditions (c) through (e) and condition (a) are transversality conditions, while conditions (g) and (h) are our maximum principles in integral form. If $\mathscr{H}$ consists of only a single point, then condition (h) becomes a trivial case.

In our second problem, we shall consider the extremal problems with restricted phase coordinates and fixed terminal time. Let us formally state our problem as follows.

PROBLEM 2. Let Q, B be defined as in Problem 1. Let $\chi_i, i = -\mu, \cdots, 0, \cdots, m$, be the continuously differentiable functions from R^{2n} into R^1, and let ϕ_i, $i = -\mu, \cdots, 0, \cdots, m$, be functionals defined on Q as follows:

$$\phi_i(x) = \chi_i(x(t_1), x(t_2)), \qquad i = -\mu, \cdots, 0, \cdots, m. \tag{2.13}$$

Let $\tilde{g}(x, t)$ be a given real-valued function defined on $G \times I$, that is twice continuously differentiable with respect to all $(x, t) \in G \times I$. Define a functional $\phi_{-\mu-1}$ on Q as follows:

$$\phi_{-\mu-1}(x) = \max_{t \in I} \tilde{g}(x(t), t). \tag{2.14}$$

Then find an element $z \in Q$ such that $\phi_i(z) \leqq 0$ for $i = -\mu - 1, -\mu, \cdots, -1$ and $\phi_i(z) = 0$ for $i = 1, \cdots, m$ and such that

$$\phi_0(z) \leqq \phi_0(x)$$

for all $x \in Q$ that satisfy the relations $\phi_i(x) \leqq 0$ for $i = -\mu - 1, -\mu, \cdots, -1$, and $\phi_i(x) = 0$ for $i = 1, \cdots, m$.

If z is a solution to Problem 2, without loss of generality, we shall assume that $\phi_{-\mu-1}(z) = 0$, and let us define an index set $\mathscr{I}_z$ and a subset I_z of I by

$$\mathscr{I}_z = \{i : i \in (-\mu, \cdots, -1), \phi_i(z) < 0\} \tag{2.15}$$

and

$$I_z = \{t : t \in I, \tilde{g}(z(t), t) < \phi_{-\mu-1}(z) = 0\}, \tag{2.16}$$

respectively.

Assuming z is a solution to Problem 2, we shall suppose that

$$\begin{gathered} \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} + \alpha' \tilde{g}_x(z(t_1), t_1) = \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} + \alpha'' \tilde{g}_x(z(t_2), t_2) = 0, \\ \alpha^i \leqq 0 \quad \text{for } i \leqq 0, \qquad \alpha^i = 0 \quad \text{for } i \in \mathscr{I}_z, \\ \alpha' = 0 \quad \text{for } t_1 \in I_z, \qquad \alpha'' = 0 \quad \text{for } t_2 \in I_z, \end{gathered} \tag{2.17}$$

imply that $\alpha' = \alpha'' = \alpha^i = 0$, $i = -\mu, \cdots, m$, where $\partial \chi_i / \partial \xi_j$, $j = 1, 2$, are defined similarly to those defined in Problem 1; and

$$\tilde{g}_x(z(t), t) \neq 0 \quad \text{for all } t \in I_z. \tag{2.18}$$

Thus we obtain the following theorem (the proof of the theorem will be given in § 4).

THEOREM 2. *If z is a solution to Problem 2 satisfying the indicated assumptions, then there exist a scalar-valued function $\lambda(t)$ defined on I, real numbers $\alpha^{-\mu}, \cdots, \alpha^0, \cdots, \alpha^m$, and an n-row vector-valued function $\bar{\psi}(t)$ defined on I such that:*

(a) *$\lambda(t)$ is nonincreasing, continuous from the right in $[t_1, t_2]$, constant on every component I_z, where $\tilde{g}(z(t), t) < 0$, and such that $\lambda(t_2) = 0$.*

(b) *$\alpha^i \leqq 0$ for $i \leqq 0$, $\alpha^i = 0$ for $i \in \{-\mu, \cdots, -1\}$ and $\phi_i(z) < 0$, and the α^i can be vanished only if $\lambda(t_1) \neq 0$.*

(c) *$\bar{\psi}(t)$ is absolutely continuous on I satisfying the following differential integral equation:*

$$\begin{aligned} \frac{d\bar{\psi}(t)}{dt} = & -\bar{\psi}(t) g^*_x(z(t), t, t) - \int_t^{t_2} [\bar{\psi}(\zeta) - \lambda(\zeta) \tilde{g}_x(z(\zeta), \zeta)] g^*_{x\zeta}(z(t), t, \zeta)\, d\zeta \\ & + \lambda(t) [\tilde{g}_{xx}(z(t), t) \dot{z}(t) + \tilde{g}_{xt}(z(t), t) + \tilde{g}_x(z(t), t) g^*_x(z(t), t, t)] \end{aligned}$$

for almost all $t \in I$.

(d) *$\bar{\psi}$ satisfies the following boundary condition:*

$$\bar{\psi}(t_1) = -\sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} + \lambda(t_1) \tilde{g}_x(z(t_1), t_1), \qquad \bar{\psi}(t_2) = \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2}.$$

(e) $$\int_{t_1}^{t_2} [\bar{\psi}(t) - \lambda(t)\tilde{g}_x(z(t), t)]\dot{h}(t)\,dt \leqq \int_{t_1}^{t_2} [\bar{\psi}(t) - \lambda(t)\tilde{g}_x(z(t), t)]\dot{h}^*(t)\,dt$$

for all $h \in \mathscr{H}$ *such that* $h(t_1) = h^*(t_1)$.

(f) $$\int_{t_1}^{t_2} [\bar{\psi}(t) - \lambda(t)\tilde{g}_x(z(t), t)]\left\{g(z(t), t, t) + \int_{t_1}^{t} g_t(z(\zeta), \zeta, t)\,d\zeta\right\} dt$$

$$\leqq \int_{t_1}^{t_2} [\bar{\psi}(t) - \lambda(t)\tilde{g}_x(z(t), t)]\left\{g^*(z(t), t, t) + \int_{t_1}^{t} g_t^*(z(\zeta), \zeta, t)\,d\zeta\right\} dt$$

for all $g \in \mathscr{G}$.

(g) $\bar{\psi}(t) - \lambda(t)\tilde{g}_x(z(t), t)$ *is not zero for* t *in some subset of* I *of positive measure.*

We remark that (f) is the maximum principle in integral form. Condition (e) is the condition which needs to be satisfied by the choice of $h^* \in \mathscr{H}$; if $\mathscr{H}$ contains only a single point, then it becomes a trivial case. Condition (c) is the adjoint system. Conditions (a), (b) and (d) serve as the transversality conditions.

3. Applications to the optimal control problems. As applications of the previous theory, we shall consider a particular class of uniformly quasi-convex functions which arise from optimal control problems with systems of Volterra nonlinear integral equations.

Let G be a nonempty open subset of R^n. Let U be an arbitrary set in R^r, and let Ω be the set of all measurable, essentially bounded functions from I into U. Let $s, t \in I$, and let $f(x, u, s, t)$ be the n-vector-valued function defined on $G \times U \times I^2$ such that:

(a) $f(x, s, s, t)$ is of class C^1 with respect to $x \in G$ for every $(u, s, t) \in U \times I^2$ and, together with $f_x(x, u, s, t)$, is continuously differentiable with respect to t for each $(x, u, s) \in G \times U \times I$, and is continuous with respect to $u \in U$ and $s \in I$ for each $(x, t) \in G \times I$;

(b) for every $u \in \Omega$, and every compact subset X of G, there exists a function $\tilde{m} \in \mathscr{L}^2$ such that

$$|f(x, u(s), s, t)| \leqq \tilde{m}(s) \quad \text{for all } (x, s, t) \in X \times I^2 \tag{3.1}$$

and such that

$$|f(x, u(s), s, t) - f(y, u(s), s, t)| \leqq \tilde{m}(s)|x - y| \tag{3.2}$$

for all $x, y \in X$, $t \in I$ and $s \in I$.

Let $\mathscr{S}$ be defined as before. We shall define the set $\mathscr{G}$ as follows: For each $u \in \Omega$, denote the function $f(x, u(s), s, t)$ from $G \times I^2$ into R^n by $g^u(x, s, t)$. It is obvious that $g^u \in \mathscr{S}$ for every $u \in \Omega$. If we denote $\mathscr{G}$ by

$$\mathscr{G} = \{g^u : u \in \Omega\}, \tag{3.3}$$

then, arguing essentially as in [2, § 4], one can show that $\mathscr{G}$ is a uniformly quasi-convex family. First, let us state the lemma which is a slight generalization of the one given by Gamkrelidze [2, § 4].

LEMMA 3.1 (Gamkrelidze). *Let* X *be a compact subset of* G. *Let* $g^{u_i}(x, s, t)$, $i = 1, \cdots, k$, *be functions from* $X \times I^2$ *into* R^n *that are measurable in* s *over* I

for each $(x, t) \in X \times I$, *of class* C^r, $r \geqq 0$, *with respect to* $x \in X$, *and such that, for some* $\bar{m}(s) \in \mathscr{L}^2(I)$,

$$|g^{u_i}(x, s, t)| \leqq \bar{m}(s), \qquad \left| \frac{\partial^j g^{u_i}}{\partial x^j}(x, s, t) \right| \leqq \bar{m}(s)$$

for all $(x, s, t) \in X \times I^2, j = 1, \cdots, r$ *and* $i = 1, \cdots, k$. *Let* $\beta^i, i = 1, \cdots, k$, *be non-negative real numbers such that* $\sum_{i=1}^{k} \beta^i = 1$. *Then, for every* $\varepsilon > 0$ *and every* $t \in I$, *it is possible to subdivide* $[t_1, t]$ *into sufficiently small mutually disjoint subintervals* $E_j, j = \pm 1, \pm 2, \cdots$, *and to assign to each* E_j *one of the functions* $g^{u_1}(x, s, t), \cdots, g^{u_k}(x, s, t)$, *which we shall denote by* g_{E_j}, *such that the function* $g(x, s, t)$, *defined by the relation*

$$g(x, s, t) = g_{E_j}(x, s, t) \quad \text{for } s \in E_j, \qquad j = \pm 1, \pm 2, \cdots,$$

satisfies the inequality

$$\left| \int_{t_1}^{\tau} \left[\sum_{i=1}^{k} \beta^i g^{u_i}(x, s, t) - g(x, s, t) \right] ds \right| < \varepsilon$$

for every $\tau \in [t_1, t]$, *and* $x \in X$.

For the proof of the lemma, please refer to [3, § 4].

Thus we conclude, through Lemma 1 and Definition 1, that the following lemma holds.

LEMMA 3.2. *If* $\mathscr{G}$ *is defined as in* (3.3), *then* $\mathscr{G}$ *is a uniformly quasi-convex family in* I.

Now, we can apply Theorems 1 and 2 to the optimal control problems with or without restricted phase coordinates. First, we shall consider the optimal control problems without restricted phase coordinates.

PROBLEM 3. Let f, $\mathscr{H}$, G, Ω be defined as above, and let χ_i, $i = -\mu, \cdots, 0, \cdots, m$, be specified as in Problem 1. Let τ be a point of I. Let x be an n-vector-valued function from I into G which satisfies the following nonlinear Volterra integral equation:

$$x(t) = h(t) + \int_{t_1}^{t} f(x(s), u(s), s, t)\, ds \qquad \text{for all } t \in [t_1, \tau], \quad \text{for some } h \in \mathscr{H} \text{ and some } u \in \Omega, \tag{3.4}$$

and such that

$$\chi_i(x(t_1), x(\tau), \tau) = 0 \quad \text{for } i = 1, \cdots, m, \tag{3.5}$$

$$\chi_i(x(t_1), x(\tau), \tau) \leqq 0 \quad \text{for } i = -\mu, \cdots, -1. \tag{3.6}$$

Then find the element (z, h^*, u^*, τ^*) which satisfies the integral equation

$$z(t) = h^*(t) + \int_{t_1}^{t} f(z(s), u^*(s), s, t)\, ds \quad \text{for all } t \in [t_1, \tau^*], \tag{3.7}$$

and the boundary conditions $\chi_i(z(t_1), z(\tau^*), \tau^*) = 0$ for $i = 1, \cdots, m$, $\chi_i(x(t_1), z(\tau^*), \tau^*) \leqq 0$ for $i = -\mu, \cdots, -1$, and such that

$$\chi_0(z(t_1), z(\tau^*), \tau^*) \leqq \chi_0(x(t_1), x(\tau), \tau) \tag{3.8}$$

for all (x, h, u, τ) satisfying (3.4)–(3.6).

It is obvious that Problem 3 is a particular case of Problem 1. Thus, we can obtain the necessary conditions for this problem similar to those in Theorem 1 except that $g(z, s, t)$ and $g^*(z, s, t)$ must be replaced by $f(z, u(s), s, t)$ and $f(z, u^*(s), s, t)$, respectively.

Next we shall consider optimal control problems with restricted phase coordinates which are a special case of Problem 2. Let us formally state the problem as follows.

PROBLEM 4. Let f, Ω, G, $\mathscr{H}$ be defined as before, and let χ_i, $i = -\mu, \cdots, 0, \cdots, m$, and $\tilde{g}$ be specified as in the statement of Problem 2. Then our problem is to find a $z \in C$ which, for some $u^* \in \Omega$ and $h^* \in \mathscr{H}$, satisfies (3.7) for all $t \in I$ and the boundary conditions

$$\chi_i(z(t_1), z(t_2)) = 0 \quad \text{for } i = 1, \cdots, m,$$

$$\chi_i(z(t_1), z(t_2)) \leqq 0 \quad \text{for } i = -\mu, \cdots, -1,$$

and $\max_{t \in I} \tilde{g}(z(t), t) \leqq 0$, and such that

$$\chi_0(z(t_1), z(t_2)) \leqq \chi_0(x(t_1), x(t_2))$$

for all $x \in C$ which, for some $u \in \Omega$ and some $h \in \mathscr{H}$, satisfies (3.4) for all $t \in I$, and such that $\chi_i(x(t_1), x(t_2)) = 0$ for $i = 1, \cdots, m$, $\chi_i(x(t_1), x(t_2)) \leqq 0$ for $i = -\mu, \cdots, -1$ and $\max_{t \in I} \tilde{g}(x(t), t) \leqq 0$.

If z is a solution to this problem, then we shall suppose that the same kind of assumptions, as given in Problem 2, hold. It is obvious that Problem 4 is a particular class of Problem 2, and we can obtain the necessary conditions for this problem similar to those of Problem 2; i.e., Theorem 2 holds with

$$g^*(x, s, t) = f(x, u^*(s), s, t). \tag{3.9}$$

4. The proof of Theorems 1 and 2. In this section we shall use the abstract multiplier rule of Neustadt [3], [4] to prove Theorems 1 and 2. Since the concept of first order convex approximation is essential in applying the multiplier rule, we shall also introduce this concept and shall construct a first order convex approximation for the set of functions which satisfy a special class of nonlinear Volterra integral equations. First, let us obtain a representation of the solution for a linear integral equation.

4.1. A linear Volterra equation. Consider the following vector-valued linear integral equation:

$$\begin{aligned} x(t) &= h(t) + \int_{t_1}^{t} K(\zeta, t)x(\zeta)\, d(\zeta) + \int_{t_1}^{t} W(\zeta, t)\, d\zeta, \qquad \zeta \leqq t, \quad t \in I, \\ x(t_1) &= h(t_1) = \xi, \end{aligned} \tag{4.1}$$

where h is a continuously differentiable vector-valued function defined on I, $K(\zeta, t)$ is a given $n \times n$ matrix-valued function defined and square integrable on $I \times I$, and $W(\zeta, t)$ is a vector-valued function defined and square integrable on $I \times I$. We assume in (4.1) that both $K(\zeta, t)$ and $W(\zeta, t)$ are continuously differentiable with respect to t for every $\zeta \in I$, and continuous in ζ over I. (For the existence and uniqueness of the solution of (4.1), please see [9, pp. 10–15].)

For each $t \in I$, let $Y(s,t)$ be the unique matrix-valued function defined for every $s \in [t_1, t]$, which, for each $t \in I$, is absolutely continuous as a function of $s \in [t_1, t]$, and is absolutely continuous as a function of $t \in [s, t_2]$, and satisfies the following differential integral equation:

$$\frac{\partial Y(s,t)}{\partial s} = -Y(s,t)K(s,s) - \int_s^t Y(\tau,t)K_\tau(s,\tau)\,d\tau \quad \text{for almost all } s \in [t_1, t], \tag{4.2}$$

together with the boundary value

$$Y(t,t) = E \quad \text{(identity matrix)}. \tag{4.3}$$

LEMMA 4.1. *Let $K(s,t)$, $W(s,t)$ and $Y(s,t)$ have the properties indicated as above. Then the solution of* (4.1) *can be represented as follows*:

$$x(t) = Y(t_1,t)\xi + \int_{t_1}^t Y(s,t)\dot{h}(s)\,ds + \int_{t_1}^t Y(s,t)\left[W(s,s) + \int_{t_1}^s W_s(\zeta,s)\,d\zeta\right]ds \tag{4.4}$$

for all $t \in I$.

Proof. After replacing t by s in (4.1), we differentiate with respect to s in both sides of (4.1). Then we obtain

$$\dot{x}(s) = K(s,s)x(s) + \dot{h}(s) + W(s,s) + \int_{t_1}^s \frac{\partial}{\partial s}K(\zeta,s)x(\zeta)\,d\zeta + \int_{t_1}^s \frac{\partial}{\partial s}W(\zeta,s)\,d\zeta.$$

If we premultiply both sides of the above equation by $Y(s,t)$ and integrate from t_1 to t, then we obtain

$$\begin{aligned}\int_{t_1}^t Y(s,t)\dot{x}(s)\,ds = &\int_{t_1}^t Y(s,t)K(s,s)x(s)\,ds + \int_{t_1}^t Y(s,t)[\dot{h}(s) + W(s,s)]\,ds \\ &+ \int_{t_1}^t\int_{t_1}^s Y(s,t)\frac{\partial}{\partial s}K(\zeta,s)x(\zeta)\,d\zeta\,ds \\ &+ \int_{t_1}^t\int_{t_1}^s Y(s,t)\frac{\partial}{\partial s}W(\zeta,s)\,d\zeta\,ds.\end{aligned}$$

Integrating by parts in the left-hand side and interchanging the order of integration in the third term of the right-hand side of the above equation, we have

$$\begin{aligned}Y(s,t)x(s)\Big|_{s=t_1}^{s=t} &- \int_{t_1}^t \frac{\partial Y(s,t)}{\partial s}x(s)\,ds \\ &= \int_{t_1}^t Y(s,t)K(s,s)x(s)\,ds + \int_{t_1}^t Y(s,t)[\dot{h}(s) + W(s,s)]\,ds \\ &\quad + \int_{t_1}^t\int_{\zeta}^t Y(s,t)\frac{\partial}{\partial s}K(\zeta,s)x(\zeta)\,ds\,d\zeta \\ &\quad + \int_{t_1}^t\int_{t_1}^s Y(s,t)\frac{\partial}{\partial s}W(\zeta,s)\,d\zeta\,ds.\end{aligned}$$

Since $Y(t, t) = E$ (identity matrix), we thus obtain (after changing s to τ, and ζ to s in the third term of the above equation)

$$x(t) = Y(t_1, t)x(t_1) + \int_{t_1}^{t} Y(s, t)[\dot{h}(s) + W(s, s)]\, ds + \int_{t_1}^{t}\int_{t_1}^{s} Y(s, t)\frac{\partial}{\partial s}W(\zeta, s)\, d\zeta\, ds$$
$$+ \int_{t_1}^{t}\left\{\frac{\partial Y(s, t)}{\partial s} + Y(s, t)K(s, s) + \int_{s}^{t} Y(\tau, t)\frac{\partial}{\partial \tau}K(s, \tau)\, d\tau\right\} x(s)\, ds. \tag{4.5}$$

Thus, (4.4) is an immediate consequence of (4.5), (4.2) and the fact that $x(t_1) = \xi$. This completes the proof of the lemma.

It is worth pointing out at this stage that, for each $s \in I$, $Y(s, t)$ also satisfies the following differential-integral equation:

$$\frac{\partial Y(s, t)}{\partial t} = K(t, t)Y(s, t) + \int_{s}^{t} \frac{\partial}{\partial t}K(\zeta, t)Y(s, \zeta)\, d\zeta \quad \text{for almost all } t \in [s, t_2]. \tag{4.6}$$

This result is not surprising since (4.2) together with the boundary condition (4.3) are the adjoint systems to the equation:

$$x(t) = \int_{t_1}^{t} K(\zeta, t)x(\zeta)\, d\zeta \tag{4.7}$$

which represents the homogeneous part of (4.1). If we replace t_1 by s, $x(t)$ by $Y(s, t)$ and differentiate with respect to t in both sides of (4.7), then we obtain (4.6). The results in (4.6) can also be obtained by postmultiplying both sides of (4.2) by $Y(v, s)$ and integrating with respect to s from v to t, and using the same arguments as in the proof of Lemma 4.1.

4.2. A first order convex approximation. Now, we shall introduce the concept of first order convex approximation of Neustadt [3].

DEFINITION 4.1 (Neustadt). Let P^v be defined as in § 2, and let Q be a set in a normed linear vector space $\mathscr{Y}$. Then a set $\mathscr{K} \subset \mathscr{Y}$ will be called a *first order convex approximation* to Q if:

(a) $0 \in \mathscr{K}$ and $\mathscr{K}$ contains points other than 0;

(b) $\mathscr{K}$ is convex;

(c) given $\{x_1, \cdots, x_v\}$, any finite subset of $\mathscr{K}$, and any $\eta > 0$, there exists a number $\varepsilon_0 > 0$ such that for every ε, $0 < \varepsilon < \varepsilon_0$, there exists a continuous map ζ_ε from P^v into Q such that the following relation holds:

$$\left\| \frac{\zeta_\varepsilon(\beta)}{\varepsilon} - \sum_{i=1}^{v} \beta^i x_i \right\| < \eta \quad \text{for all } \beta = (\beta^1, \cdots, \beta^v) \in P^v. \tag{4.8}$$

Let $\mathscr{H}$, $\mathscr{S}$, C and G be defined as in § 2, and let $\mathscr{G}$ be a uniformly quasi-convex set in $\mathscr{S}$. Let Q be the set of functions $x(t)$ defined on I and taking values on G which satisfy the nonlinear Volterra equation (2.8) for some $h \in \mathscr{H}$ and $g \in \mathscr{G}$, and also satisfy the initial condition $x(t_1) = \xi$ for some $\xi \in R^n$.

Let z be the element of C taking values on G such that it satisfies (2.10) for some $h^* \in \mathscr{H}$ and some $g^* \in \mathscr{G}$, and the boundary condition (2.11) for some $\xi^* \in R^n$. It is obvious that $z \in Q$. The elements of $\mathscr{H} - h^*$ will be denoted by δh and those of $[\mathscr{G}] - g^*$ by δg.

Let $Y(s, t)$ be the matrix-valued function described in the previous section such that, for each $t \in I$,

$$\frac{\partial Y(s,t)}{\partial s} = -Y(s,t)g_x^*(z(s),s,s) - \int_s^t Y(\tau,t)g_{x\tau}^*(z(s),s,\tau)\,d\tau \tag{4.9}$$

for almost all $s \in [t_1, t]$,

and such that, for each $s \in I$,

$$\frac{\partial Y(s,t)}{\partial t} = g_x^*(z(t),t,t)Y(s,t) + \int_s^t g_{xt}^*(z(\tau),\tau,t)Y(s,\tau)\,d\tau \tag{4.10}$$

for almost all $t \in [s, t_2]$,

and such that (4.3) holds.

For each $\xi \in R^n$, each $\delta h \in \mathscr{H} - h^*$ and each $\delta g \in [\mathscr{G}] - g^*$, let us define $\delta x(\cdot\,; \xi, \delta h, \delta g)$ in C as follows:

$$\begin{aligned}\delta x(t;\xi,\delta h,\delta g) &= Y(t_1,t)\xi + \int_{t_1}^t Y(s,t)\dot{\delta h}(s)\,ds \\ &\quad + \int_{t_1}^t Y(s,t)\left[\delta g(z(s),s,s) + \int_{t_1}^s \delta g_s(z(\zeta),\zeta,s)\,d\zeta\right]ds\end{aligned} \tag{4.11}$$

for all $t \in I$.

Let $\mathscr{K}$ be the following subset of the space C:

$$\mathscr{K} = \{\delta x(\cdot\,; \xi, \delta h, \delta g);\ \xi \in R^n, \delta h \in \mathscr{H} - h^*, \delta g \in [\mathscr{G}] - g^*\}. \tag{4.12}$$

Now, let G be an open subset of R^n, and X be a compact subset of G. Let $\mathscr{F}$ be a given family of equicontinuous functions from I into X. We have the following lemma which is similar to [4, Lemma 3.1] or [10, Lemma 2.1].

LEMMA 4.2. *Let $g \in \mathscr{S}$, and $\tilde{m}$ be some nonnegative integrable function on I, and $\mathscr{F}$ be a preassigned set defined as above. Then, for every $\gamma > 0$, there exists an $\eta > 0$ such that, for every $t \in I$, if*

$$|g(x,s,t) - g(y,s,t)| \leqq \tilde{m}(s)|x - y| \quad \textit{for all } x, y \in X \quad \textit{and all } s \in [t_1, t], \tag{4.13}$$

$$\left|\int_{t_1}^{\tau'} g(x,s,t)\,ds\right| < \eta \quad \textit{for all } x \in X \quad \textit{and} \quad \tau' \in [t_1, t], \tag{4.14}$$

then

$$\left|\int_{t_1}^{\tau''} g(x(s),s,t)\,ds\right| < \gamma \quad \textit{for all } \tau'' \in [t_1, t] \quad \textit{and all possible } x \textit{ of } \mathscr{F}. \tag{4.15}$$

Proof. Let $\tilde{m}$ and $\gamma > 0$ be given as in the lemma statement. Since $\mathscr{F}$ is equicontinuous in I, we can subdivide I in such a way that if $t_1 = s_0 < s_1 < \cdots < s_l = t_2$, then

$$|x(s) - x(s_j)| < \gamma\left(2\int_I \tilde{m}(s)\,ds\right)^{-1} \tag{4.16}$$

for all $s \in [s_j, s_{j+1}]$, $j = 1, \cdots, l - 1$, and all $x \in \mathscr{F}$.

Now let us pick an arbitrary t of I and fix this t for the remainder of our arguments.

By virtue of (4.14), it is obvious that

$$\left|\int_{s_j}^{s'} g(x(s_j), s, t)\, ds\right| < 2\eta \quad \text{for all } s' \in [s_j, s_{j+1}] \text{ and } s' \leqq t, \quad j = 1, \cdots, l-1. \tag{4.17}$$

Let us put $\eta = \gamma/4l$ and estimate the quantity in the left-hand side of (4.15). Let us suppose that, without loss of generality, $\tau'' = s_{j'+1}$ for some $j' \in \{0, 1, \cdots, l-1\}$ and $\tau'' \leqq t$. Thus, it follows from (4.13), (4.16) and (4.17) that

$$\begin{aligned}
\left|\int_{t_1}^{\tau''} g(x(s), s, t)\, ds\right| &= \left|\sum_{j=0}^{j'} \int_{s_j}^{s_{j+1}} g(x(s), s, t)\, ds\right| \\
&\leqq \sum_{j=0}^{j'} \int_{s_j}^{s_{j+1}} [|g(x(s_j), s, t)| + |g(x(s), s, t) - g(x(s_j), s, t)|]\, ds \\
&\leqq 2l\eta + \sum_{j=0}^{j'} \int_{s_j}^{s_{j+1}} \tilde{m}(s)|x(s) - x(s_j)|\, ds \\
&< 2l\eta + \sum_{j=0}^{l-1} \int_{s_j}^{s_{j+1}} \tilde{m}(s)\gamma\left(2\int_I \tilde{m}(s)\, ds\right)^{-1} ds \\
&\leqq 2l\eta + \gamma/2.
\end{aligned}$$

Since $\eta = \gamma/4l$, (4.15) follows immediately. This completes the proof of the lemma.

Thus, by virtue of Lemma 4.1, Lemma 4.2 and Definition 4.1, arguing as in [4], we can prove the following important lemma.

LEMMA 4.3. *Let Q, z, $\mathscr{K}$ be defined as before, and let $Q^* = Q - z$. Then $\mathscr{K}$ is a first order convex approximation to Q^* in the space C.*

Proof. By the definition of a first order convex approximation, we have to prove that:

(a) $0 \in \mathscr{K}$, and $\mathscr{K}$ contains points other than zero;

(b) $\mathscr{K}$ is convex;

(c) given $\{\delta x_1, \cdots, \delta x_\nu\}$, any finite subset of $\mathscr{K}$ and any $\eta > 0$, there exists a number $\varepsilon_0 > 0$ such that, for every ε, $0 < \varepsilon < \varepsilon_0$, there exists a continuous map ζ_ε from P^ν into Q^* such that relation (4.8) holds with x_i replaced by δx_i.

By (4.11), $\delta x(\cdot\,; 0, 0, 0) = 0$; it then follows from the definition of $\mathscr{K}$ that $0 \in \mathscr{K}$. It is also clear that $\mathscr{K}$ contains points other than zero. The convexity property of $\mathscr{K}$ follows from the fact that

$$\sum_{i=1}^{\nu} \beta^i\, \delta x(\cdot\,; \xi_i, \delta h_i, \delta g_i) = \delta x\left(\cdot\,; \sum_{i=1}^{\nu} \beta^i \xi_i, \sum_{i=1}^{\nu} \beta^i \delta h_i, \sum_{i=1}^{\nu} \beta^i \delta g_i\right)$$

for all $\beta = (\beta^1, \cdots, \beta^\nu) \in P^\nu$.

Now, we want to prove that given any finite subset $\{\delta x_1, \cdots, \delta x_2\}$ of $\mathscr{K}$ and any $\eta > 0$ there exists a number $\varepsilon_0 > 0$ such that, for every ε, $0 < \varepsilon \leqq \varepsilon_0$, there exists a continuous map ζ_ε from P^ν into Q^* such that relation (4.8) holds with x_i replaced by δx_i.

Let $\{\delta x_1, \cdots, \delta x_\nu\}$ be any finite subset of $\mathscr{K}$, and $\beta = (\beta^1, \cdots, \beta^\nu)$ be any vector in P^ν. Let $\sum_{i=1}^{\nu} \beta^i \delta x_i$ be denoted by $\delta x(\cdot\,; \beta)$. For every δx_i, there exist $\xi_i \in R^n$, $\delta h_i \in \mathscr{H} - h^*$ and $\delta g_i \in [\mathscr{G}] - g^*$ such that $\delta x_i = \delta x(\cdot\,; \xi_i, \delta h_i, \delta g_i)$, $i = 1, \cdots, \nu$. For each $\beta \in P^\nu$, let $\sum_{i=1}^{\nu} \beta^i \xi_i = \xi_\beta$, $\sum_{i=1}^{\nu} \beta^i \delta h_i = \delta h(\cdot\,; \beta)$ and

$$\sum_{i=1}^{\nu} \beta^i \delta g_i(x, s, t) = \delta g(x, s, t; \beta). \tag{4.18}$$

It then follows from (4.9), (4.11), (4.12) and Lemma 4.1 that

$$\delta x(t; \beta) = \delta h(t; \beta) + \int_{t_1}^{t} g_x^*(z(s), s, t)\, \delta x(s; \beta)\, ds + \int_{t_1}^{t} \delta g(z(s), s, t; \beta)\, ds \tag{4.19}$$

for all $t \in I$ and every $\beta \in P^\nu$, and that

$$\delta x(t_1; \beta) = \xi_\beta. \tag{4.20}$$

Since $\delta g_i \in [\mathscr{G}] - g^*$, there are functions $g_1, \cdots, g_{\nu'} \in \mathscr{G}$ and vectors $\beta_i = (\beta_i^1, \cdots, \beta_i^{\nu'}) \in P^{\nu'}$ such that

$$\delta g_i = \sum_{j=1}^{\nu'} \beta_i^j g_j - g^*, \qquad i = 1, \cdots, \nu, \tag{4.21}$$

Now let X be a compact subset of G such that $z(t)$ is an interior point of X for every $t \in I$.

Note that for each g_j and g^* of $\mathscr{G}$, there exists a function $m_j \in \mathscr{L}^2[I]$, $j = 0, \cdots, \nu$, such that

$$|g_j(x, s, t)| \leqq m_j(s), \tag{4.22}$$

$$|g^*(x, s, t)| \leqq m_0(s), \tag{4.23}$$

$$|g_j(x, s, t) - g_j(y, s, t)| \leqq m_j(s)\, |x - y|, \tag{4.24}$$

$$|g^*(x, s, t) - g^*(y, s, t)| \leqq m_0(s)\, |x - y| \tag{4.25}$$

for every $x, y \in X$ and every $(s, t) \in I \times I$. By the definition of uniform quasi-convexity, (4.18), (4.21)–(4.25), it then follows that there exists a function $\bar{m}_1 \in \mathscr{L}^2[I]$ such that

$$|\delta g(x, s, t; \beta)| \leqq \bar{m}_1(s), \tag{4.26}$$

$$|\delta g(x, s, t; \beta) - \delta g(y, s, t; \beta)| \leqq \bar{m}_1(s)\, |x - y| \tag{4.27}$$

for every $\beta \in P^\nu$, all $x, y \in X$ and every $(s, t) \in I \times I$.

Note that for each $\beta \in P^\nu$ and every ε, $0 < \varepsilon \leqq 1$,

$$\begin{aligned} g^*(x, s, t) + \varepsilon\, \delta g(x, s, t; \beta) &= g^*(x, s, t) + \varepsilon \sum_{i=1}^{\nu} \sum_{j=1}^{\nu'} \beta^i \beta_i^j g_j(x, s, t) - \varepsilon g^*(x, s, t) \\ &= (1 - \varepsilon) g^*(x, s, t) + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu'} \varepsilon \beta^i \beta_i^j g_j(x, s, t). \end{aligned}$$

Since $(1 - \varepsilon) + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu'} \varepsilon \beta^i \beta_i^j = 1$, we conclude that

$$g^*(x, s, t) + \varepsilon\, \delta g(x, s, t; \beta) \in [\mathscr{G}], \qquad 0 < \varepsilon \leqq 1. \tag{4.28}$$

If $\eta > 0$ is an arbitrary number and $\varepsilon \in (0, 1]$ is a fixed number, then, by Definition 2.1 and (4.28), there exist functions $g_\eta(x, s, t; \beta) \in \mathscr{G}$, defined for every $\beta \in P^\nu$, such that the functions

$$\delta^2 g_\eta(x, s, t; \beta) = g_\eta(x, s, t; \beta) - [g^*(x, s, t) + \varepsilon\, \delta g(x, s, t; \beta)] \tag{4.29}$$

satisfy the following conditions:

$$|\delta^2 g_\eta(x, s, t; \beta)| \leqq \tilde{m}(s) \tag{4.30}$$

and

$$|\delta^2 g_\eta(x, s, t; \beta) - \delta^2 g_\eta(y, s, t; \beta)| \leqq \tilde{m}(s)|x - y| \tag{4.31}$$

for every $(x, s, t) \in X \times I \times I$, every $(y, s, t) \in X \times I \times I$, every $\beta \in P^\nu$ and for some $\tilde{m} \in \mathscr{L}^2$;

$$\left| \int_{t_1}^{\tau} \delta^2 g_\eta(x, s, t; \beta)\, ds \right| < \eta \tag{4.32}$$

for every $(x, t) \in X \times I$, $\tau \in I$ and every $\beta \in P^\nu$;

$$\sup_{\bar{t} \in I} \left| \int_{t_1}^{\bar{t}} [\delta^2 g_\eta(x, s, t; \beta_i) - \delta^2 g_\eta(x, s, t; \beta)]\, ds \right| \xrightarrow[\substack{\beta_i \to \beta \\ \{\beta_i\} \in P^\nu}]{} 0 \tag{4.33}$$

for every $(x, t) \in X \times I$ and $\beta \in P^\nu$. It then follows from Lemma 4.2 and (4.28)–(4.32) that there exists a number $\eta_0 > 0$ such that

$$\left| \int_{t_1}^{\tau} \delta^2 g_{\eta_0}(x(s), s, t; \beta)\, ds \right| < \varepsilon^2 \tag{4.34}$$

for all $\tau \in [t_1, t]$ for all possible x of equicontinuous functions from I into X. Since the number $\varepsilon \in (0, 1]$ is arbitrary, it is evident that the functions $g_{\eta_0}(x, s, t; \beta)$ and $\delta^2 g_{\eta_0}(x, s, t; \beta)$, as well as η_0 itself, depend on ε. Therefore, we shall write, for every $\beta \in P^\nu$,

$$g_{\eta_0}(x, s, t; \beta) = g(x, s, t; \beta, \varepsilon), \tag{4.35}$$

$$\delta^2 g_{\eta_0}(x, s, t; \beta) = \delta^2 g(x, s, t; \beta, \varepsilon). \tag{4.36}$$

It then follows from (4.22)–(4.25), (4.26), (4.27), (4.29)–(4.30) that if $\hat{m}(s) = m_0(s) + \bar{m}(s) + \tilde{m}(s)$, then

$$|g(x, s, t; \beta, \varepsilon)| \leqq \hat{m}(s) \tag{4.37}$$

and

$$|g(x, s, t; \beta, \varepsilon) - g(y, s, t; \beta, \varepsilon)| \leqq \hat{m}(s)\,|x - y| \tag{4.38}$$

for every $(x, s, t) \in X \times I \times I$, every $\varepsilon \in (0, 1]$ and every $\beta \in P^\nu$.

For every $\varepsilon \in (0, 1]$ and every $\beta \in P^\nu$, let us consider the perturbation equation

$$\begin{aligned} x(t; \beta, \varepsilon) &= h(t; \beta, \varepsilon) + \int_{t_1}^{t} g(x(s; \beta, \varepsilon), s, t; \beta, \varepsilon)\, dt \\ &= h(t; \beta, \varepsilon) + \int_{t_1}^{t} [g^*(x(s; \beta, \varepsilon), s, t) + \varepsilon\, \delta g(x(s; \beta, \varepsilon), s, t; \beta) \\ &\qquad + \delta^2 g(x(s; \beta, \varepsilon), s, t; \beta, \varepsilon)]\, ds \end{aligned} \tag{4.39}$$

for all $t \in I$, and $x(t_1; \beta, \varepsilon) = h(t_1; \beta, \varepsilon)$, where $h(t; \beta, \varepsilon) = h^*(t) + \varepsilon\, \delta h(t; \beta) = h^*(t) + \varepsilon \sum_{i=1}^{\nu} \beta^i\, \delta h_i(t)$. (For the existence and uniqueness of solutions of nonlinear Volterra integral equations of type (4.39) please see [9, pp. 42–47].)

Now we shall show that if $x(t; \beta, \varepsilon)$ is the solution of (4.39) then there exists a number ε_1 such that, if $0 < \varepsilon \leqq \varepsilon_1$, then $x(t; \beta, \varepsilon)$ is defined on X for all $t \in I$ and every $\beta \in P^{\nu}$, such that

$$\lim_{\varepsilon \to 0^+} \max_{t \in I} |x(t; \beta, \varepsilon) - z(t)| = 0. \tag{4.40}$$

By virtue of (2.11) and (4.29), we obtain

$$\begin{aligned} x(t; \beta, \varepsilon) - z(t) = \varepsilon & \left[\delta h(t; \beta) + \int_{t_1}^{t} \delta g(x(s; \beta, \varepsilon), s, t; \beta)\, ds \right] \\ & + \int_{t_1}^{t} \delta^2 g(x(s; \beta, \varepsilon), s, t; \beta, \varepsilon)\, ds \\ & + \int_{t_1}^{t} [g^*(x(s; \beta, \varepsilon), s, t) - g^*(z(s), s, t)]\, ds. \end{aligned} \tag{4.41}$$

It then follows from (4.25), (4.26), (4.34) and (4.36) that

$$\begin{aligned} |x(t; \beta, \varepsilon) - z(t)| &\leqq \varepsilon \left(\varepsilon + |\delta h(t; \beta)| + \int_{t_1}^{t} \bar{m}_1(s)\, ds \right) + \int_{t_1}^{t} m_0(s)\, |x(s; \beta, \varepsilon) - z(s)|\, ds \\ &\leqq \varepsilon \left(1 + \max_{t \in I} |\delta h(t; \beta)| + \int_{t_1}^{t} \bar{m}_1(s)\, ds \right) \\ &\quad + \int_{t_1}^{t} m_0(s)\, |x(s; \beta, \varepsilon) - z(s)|\, ds. \end{aligned}$$

If we let

$$\eta = \left(1 + \max_{t \in I} |\delta h(t; \beta)| + \int_{I} \bar{m}_1(s)\, ds \right) \exp \left(\int_{I} m_0(s)\, ds \right),$$

then it follows from the Gronwall inequality that

$$|x(t; \beta, \varepsilon) - z(t)| \leqq \varepsilon \eta \tag{4.42}$$

for all $t \in I$, every $\beta \in P^{\nu}$ and all $\varepsilon \in (0, 1]$. Let $\bar{\eta}$ be the distance between $z(t)$ and $G - X$, and let $\varepsilon_1 = \min(1, \bar{\eta}/2\eta)$. Then we see that if $0 < \varepsilon \leqq \varepsilon_1$, then $x(t; \beta, \varepsilon)$ is defined on X, and furthermore, (4.40) is valid.

For every $\varepsilon \in (0, \varepsilon_1]$, every $t \in I$ and every $\beta \in P^{\nu}$, we obtain from (4.41) that

$$\begin{aligned} \frac{x(t; \beta, \varepsilon) - z(t)}{\varepsilon} &= \delta h(t; \beta) + \int_{t_1}^{t} \delta g(z(s), s, t)\, ds \\ &\quad + \frac{1}{\varepsilon} \int_{t_1}^{t} g_x^*(z(s), s, t)[x(s; \beta, \varepsilon) - z(s)]\, ds + \lambda(t; \beta, \varepsilon), \end{aligned} \tag{4.43}$$

where

$$\begin{aligned}\lambda(t;\beta,\varepsilon) = &\int_{t_1}^{t} [\delta g(x(s;\beta,\varepsilon),s,t;\beta) - \delta g(z(s),s,t)]\,ds \\ &+ \frac{1}{\varepsilon}\int_{t_1}^{t} \{[g^*(x(s;\beta,\varepsilon),s,t) - g^*(z(s),s,t)] \\ &\qquad - g_x^*(z(s),s,t)[x(s;\beta,\varepsilon) - z(s)]\}\,ds \\ &+ \frac{1}{\varepsilon}\int_{t_1}^{t} \delta^2 g(x(s;\beta,\varepsilon),s,t;\beta,\varepsilon)\,ds.\end{aligned} \tag{4.44}$$

Arguing as in [4], one can show without difficulty that if

$$\lambda(\varepsilon) = \max_{(t,\beta)\in I\times P^\nu} |\lambda(t;\beta,\varepsilon)|,$$

then

$$\lim_{\varepsilon\to 0^+} \lambda(\varepsilon) = 0. \tag{4.45}$$

Now it follows from (4.19), (4.25) and (4.44) that

$$\left|\frac{x(t;\beta,\varepsilon) - z(t)}{\varepsilon} - \delta x(t;\beta)\right| \leqq \lambda(\varepsilon) + \int_{t_1}^{t} m_0(s)\left|\frac{x(t;\beta,\varepsilon) - z(s)}{\varepsilon} - \delta x(s;\beta)\right| ds.$$

By (4.45) and the Gronwall inequality, we conclude immediately that

$$\lim_{\varepsilon\to 0^+}\left|\frac{x(t;\beta,\varepsilon) - z(t)}{\varepsilon} - \delta x(t;\beta)\right| = 0$$

uniformly with respect to $(t,\beta)\in I\times P^\nu$. In particular, we obtain

$$\lim_{\varepsilon\to 0^+}\max_{t\in I}\left|\frac{x(t;\beta,\varepsilon) - z(t)}{\varepsilon} - \delta x(t;\beta)\right| = 0 \tag{4.46}$$

uniformly in $\beta\in P^\nu$. For every $\varepsilon\in(0,\varepsilon_1]$ and every $\beta\in P^\nu$, it can also be proved that

$$\lim_{\substack{\beta'\to\beta\\ \beta'\in P^\nu}} |x(t;\beta',\varepsilon) - x(t;\beta,\varepsilon)| = 0. \tag{4.47}$$

Now if we let $\zeta_\varepsilon(\beta) = x(t;\beta,\varepsilon) - z(t)$, for every $\varepsilon\in(0,\varepsilon_1]$, it then follows from (4.46)–(4.47) that K is a first order convex approximation to Q^*. This completes the proof of the lemma.

4.3. The proof of Theorem 1. It can be easily seen that Problem 1 falls under the category of the canonical optimization problem in the sense of [3, § 4], with $W = Q\times I$ and $Q' = Q\times B$. If ϕ_i, $i = -\mu,\cdots,0,\cdots,m$, are the functionals defined in the statement of the problem and if (z,τ^*) is a solution to this problem, then it is known (see [4]) that

$$(\phi_i(z+\varepsilon y,\tau^*+\varepsilon\Delta') - \phi_i(z,\tau^*))/\varepsilon \xrightarrow[\substack{\varepsilon\to 0\\ y\to\delta x\\ \Delta'\to\Delta}]{} l_i(\delta x,\Delta) \tag{4.48}$$

for $i = -\mu, \cdots, 0, \cdots, m$, where l_i is the linear functional on $C \times R^1$ given by

$$l_i(\delta x, \Delta) = \frac{\partial \chi_i}{\partial \xi_1} \delta x(t_1) + \frac{\partial \chi_i}{\partial \xi_2}[\delta x(\tau^*) + \dot{z}(\tau^*)\Delta] + \frac{\partial \chi_i}{\partial \tau}\Delta, \tag{4.49}$$

and $y \to \delta x$ indicates convergence in the norm topology of C.

It can also been seen that if (z, τ^*) is a solution to Problem 1 satisfying the indicated hypothesis such that (4.48) and (4.49) are satisfied, then (z, τ^*) is a smoothly regular solution of our problem in the sense of [3, Definition 4.4].

By virtue of Lemma 4.3, we know that $\mathscr{K}$ is a first order convex approximation to $Q - z$ in the space C. Since B is a convex subset of I, $B - \tau^*$ is a first order convex approximation to itself. Hence, $\mathscr{K} \times [B - \tau^*]$ is a first order convex approximation to $[Q - z] \times [B - \tau^*] = Q \times B - (z, \tau^*)$ (see [3, Notes 2.2 and 2.3]). It now follows from [3, Theorem 4.2], that there exist real numbers $\alpha^{-\mu}, \cdots, \alpha^0, \cdots, \alpha^m$ not all zero such that

$$\alpha^i \leqq 0 \quad \text{for } i \leqq 0 \quad \text{and} \quad \alpha^i \phi_i(z, \tau^*) = 0 \quad \text{if } i < 0, \tag{4.50}$$

$$\sum_{i=-\mu}^{m} \alpha^i l_i(\delta x, \Delta) \leqq 0 \quad \text{for all } (\delta x, \Delta) \in \mathscr{K} \times (B - \tau^*), \tag{4.51}$$

$$\sum_{i=-\mu}^{m} \alpha^i l_i \neq 0. \tag{4.52}$$

Since $[\delta x(\,\cdot\,; \xi, \delta h, \delta g), 0] \in \mathscr{K} \times (B - \tau^*)$ for all $\delta h \in \mathscr{H} - h^*$ and all $\delta g \in [\mathscr{G}] - g^*$, it follows from (4.51) and (4.49) that

$$\sum_{i=-\mu}^{m} \alpha^i \left\{ \frac{\partial \chi_i}{\partial \xi_1} \delta x(t_1) + \frac{\partial \chi_i}{\delta \xi_2} \delta x(\tau^*) \right\} \leqq 0 \quad \text{for all } \delta x(\,\cdot\,; \xi, \delta h, \delta g) \in \mathscr{K}. \tag{4.53}$$

By virtue of the fact that $\delta x(\,\cdot\,; 0, 0, \delta g) \in \mathscr{K}$, we obtain, through (4.11), (4.12) and (4.53), that

$$\sum_{i=-\mu}^{m} \alpha^i \left\{ \frac{\partial \chi_i}{\partial \xi_2} \cdot \int_{t_1}^{\tau^*} Y(s, \tau^*) \left[\delta g(z(s), s, s) + \int_{t_1}^{s} \delta g_s(z(\zeta), \zeta, s)\, d\zeta \right] ds \right\} \leqq 0 \tag{4.54}$$

for all $\delta g \in [\mathscr{G}] - g^*$.

If we let

$$\psi(s) = \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} Y(s, \tau^*) \quad \text{for all } s \in I, \tag{4.55}$$

and use the fact that $\delta g = g - g^*$ for all $g \in [\mathscr{G}]$, then (4.54) can be rewritten as

$$\begin{aligned} &\int_{t_1}^{\tau^*} \psi(s) \left[g(z(s), s, s) + \int_{t_1}^{s} g_s(z(\zeta), \zeta, s)\, d\zeta \right] ds \\ &\qquad \leqq \int_{t_1}^{\tau^*} \psi(s) \left[g^*(z(s), s, s) + \int_{t_1}^{s} g_s^*(z(\zeta), \zeta, s)\, d\zeta \right] ds \quad \text{for all } g \in \mathscr{G}. \end{aligned} \tag{4.56}$$

This is our maximum principle, i.e., condition (h) in Theorem 1.

If we pick $[\delta x(\cdot\,; 0, 0, 0), \Delta] \in \mathscr{K} \times (B - \tau^*)$, then it follows from (4.51) and (4.49) that

$$\sum_{i=-\mu}^{m} \alpha^i \left\{ \frac{\partial \chi_i}{\partial \xi_2} \dot{z}(\tau^*) + \frac{\partial \chi_i}{\partial \tau} \right\} \cdot \Delta \leqq 0 \quad \text{for all } \Delta \in B - \tau^*.$$

By assumption, τ^* is an interior point of B. Hence the above inequality is possible only if

$$\sum_{i=-\mu}^{m} \alpha^i \left\{ \frac{\partial \chi_i}{\partial \xi_2} \cdot \dot{z}(\tau^*) + \frac{\partial \chi_i}{\partial \tau} \right\} = 0. \tag{4.57}$$

Since $Y(\tau^*, \tau^*) = E$ (identity matrix), it follows from (4.55) that

$$\psi(\tau^*) = \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2}. \tag{4.58}$$

Thus (4.57) implies that

$$\psi(\tau^*) \cdot \dot{z}(\tau^*) = -\sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \tau}. \tag{4.59}$$

If we choose $\delta x(\cdot\,; \xi, \xi, 0) \in \mathscr{K}$, then it follows from (4.53), (4.11) and (4.12) that

$$\sum_{i=-\mu}^{m} \alpha^i \left\{ \frac{\partial \chi_i}{\partial \xi_1} \xi + \frac{\partial \chi_i}{\partial \xi_2} \cdot Y(t_1, \tau^*) \xi \right\} \leqq 0 \quad \text{for all } \xi \in R^n. \tag{4.60}$$

This is possible only if

$$\sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} Y(t_1, \tau^*) = 0.$$

By virtue of (4.55), we thus have that

$$\psi(t_1) = -\sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1}. \tag{4.61}$$

Since $\delta x(\cdot\,; 0, \delta h, 0)$ also belongs to $\mathscr{K}$, it follows immediately from (4.53), (4.11) and (4.12) that

$$\sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} \int_{t_1}^{\tau^*} Y(s, \tau^*)\, \dot{\delta h}(s)\, ds \leqq 0 \quad \text{for all } \delta h \in \mathscr{H} - h^*$$

such that $\delta h(t_1) = 0$, or, equivalently, by virtue of (4.55), it can be put in the following form:

$$\int_{t_1}^{\tau^*} \psi(s) \dot{h}(s)\, ds \leqq \int_{t_1}^{\tau^*} \psi(s) \dot{h}^*(s)\, ds \quad \text{for all } h \in \mathscr{H} \tag{4.62}$$

such that $h(t_1) = h^*(t_1)$.

Thus we have proved Theorem 1 except for conditions (b) and (f). First, let us observe that condition (b) is the immediate consequence of (4.9) and (4.55). So the only thing left to be proved is condition (f). It is clear that, by the definition of ψ and $Y(s, t)$, ψ is absolutely continuous in I. But, by the assumption, we know that $\psi(\tau^*) \neq 0$ and $\psi(t_1) \neq 0$, so through the continuity argument, we conclude that $\psi(s)$ is not equal to 0 in a subset of I of positive measure. This completes the proof of Theorem 1.

4.4. The proof of Theorem 2. In this section, we shall consider Problem 2 and shall prove the results in Theorem 2.

Let Q, B, χ_i, $i = -\mu, \cdots, 0, \cdots, m$, ϕ_i, $i = -\mu - 1, -\mu, \cdots, 0, \cdots, m$, $\mathscr{I}_z$ and I_z be defined as in the statements of Problem 2. Let the assumptions in Problem 2 also hold. Hence, arguing as in [4] (see [4, § 5]), we can show that there exists a convex, continuous functional $l_{-\mu-1}$ on C such that

$$\frac{\phi_{-\mu-1}(z + \varepsilon\delta y) - \phi_{-\mu-1}(z)}{\varepsilon} \xrightarrow[\delta y \to \delta x]{\varepsilon \to 0^+} l_{-\mu-1}(\delta x) \quad \text{for all } \delta x \in C, \tag{4.63}$$

where

$$l_{-\mu-1}(\delta x) = \sup_{\substack{t \notin I_z \\ t \in I}} [\tilde{g}_x(z(t), t)\, \delta x(t)]. \tag{4.64}$$

It was also shown in [4, § 5] that there are linear functionals $l_i, i = -\mu, \cdots, 0, \cdots, m$, on C such that

$$\frac{\phi_i(z + \varepsilon\, \delta y) - \phi_i(z)}{\varepsilon} \xrightarrow[\delta y \to \delta x]{\varepsilon \to 0} l_i(\delta x), \tag{4.65}$$

where

$$l_i(\delta x) = \frac{\partial \chi_i}{\partial \xi_1}(z(t_1), z(t_2))\, \delta x(t_1) + \frac{\partial \chi_i}{\partial \xi_2}(z(t_1), z(t_2))\, \delta x(t_2), \tag{4.66}$$

$$i = -\mu, \cdots, 0, \cdots, m.$$

Note that $\delta y \to \delta x$ in (4.63) and (4.65) denotes convergence in the norm topology of C. It can easily be seen that Problem 2 is a special case of the canonical optimization problem defined in [3, § 4] with $W = Q' = Q$. If z is a solution to this problem satisfying (2.10) and (2.11) for some $g^* \in \mathscr{G}$, some $h^* \in \mathscr{H}$ and some $\xi^* \in R^n$, then, by virtue of (4.63)–(4.66) and our hypothesis, z is a totally regular solution of Problem 2 in the sense of [3, Definition 4.3]. Let us define a cone $Z_{-\mu-1}$ in C as follows:

$$Z_{-\mu-1} = \{\delta y : \delta y \in C, \sup_{\substack{t \notin I_z \\ t \in I}} [\tilde{g}_x(z(t), t)\, \delta y(t)] < 0\} \cup \{0\}. \tag{4.67}$$

Thus, on the basis of [3, Lemma 4.4], z is also a regular solution to Problem 2.

Now, since $\mathscr{K}$ is a first order convex approximation to $Q - z$ and z is a regular solution to Problem 2, we conclude, through [3, Theorem 4.1], that there exist real numbers α^i, $i = -\mu, \cdots, 0, \cdots, m$, and a functional $\bar{l}_{-\mu-1} \in C^*$, the dual space of C, such that

$$\bar{l}_{-\mu-1}(\delta x) + \sum_{i=-\mu}^{m} \alpha^i l_i(\delta x) \leqq 0 \quad \text{for all } \delta x \in \mathscr{K}, \tag{4.68}$$

$$\alpha^i \leqq 0 \quad \text{for } i \leqq 0, \qquad \alpha^i = 0 \quad \text{for } i \in \mathscr{I}_z,$$

$$\bar{l}_{-\mu-1}(\delta y) \geqq 0 \quad \text{for all } \delta y \in Z_{-\mu-1}, \tag{4.69}$$

and

$$\bar{l}_{-\mu-1} + \sum_{i=-\mu}^{m} \alpha^i l_i \neq 0. \tag{4.70}$$

Let us embed C into the Banach space $\mathscr{B}_1$ of all continuous functions from I into R^n with the sup norm, and extend $\bar{l}_{-\mu-1}$ onto $\mathscr{B}_1$. Thus, following the same procedure as in [4, p. 8], we can show that, through hypothesis (2.18) and [4, Lemma 8.1], there exists a scalar-valued, nonincreasing function $\lambda(t)$, defined on I and continuous from the right in (t_1, t_2), such that

$$\lambda(t_2) = 0, \qquad \lambda(t) \text{ is constant on every component of } I_z, \tag{4.71}$$

and

$$\bar{l}_{-\mu-1}(\delta y) = \int_{t_1}^{t_2} \tilde{g}_x(z(t), t)\, \delta y(t)\, d\lambda(t) \quad \text{for all } \delta y \in C. \tag{4.72}$$

Thus, it follows from (4.68), (4.72), (4.66) and (4.12) that

$$\begin{aligned} &\int_{t_1}^{t_2} \tilde{g}_x(z(t), t)\, \delta x(t; \xi, \delta h, \delta g)\, d\lambda(t) + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} \xi \\ &\quad + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} \cdot \delta x(t_2; \xi, \delta h, \delta g) \leqq 0 \end{aligned} \tag{4.73}$$

$$\text{for all } \delta x(\cdot\,; \xi, \delta h, \delta g) \in \mathscr{K}.$$

Since $\delta x(\cdot\,; 0, 0, \delta g) \in \mathscr{K}$, relations (4.73) and (4.11) imply that

$$\begin{aligned} &\int_{t_1}^{t_2} \tilde{g}_x(z(t), t) \int_{t_1}^{t} Y(s, t) \left[\delta g(z(s), s, s) + \int_{t_1}^{s} \delta g_s(z(\zeta), \zeta, s)\, d\zeta \right] ds\, d\lambda(t) \\ &+ \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} \cdot \int_{t_1}^{t_2} Y(s, t_2) \left[\delta g(z(s), s, s) + \int_{t_1}^{s} \delta g_s(z(\zeta), \zeta, s)\, d\zeta \right] ds \leqq 0 \end{aligned} \tag{4.74}$$

$$\text{for all } t \in I, \quad \text{and for all } \delta g \in [\mathscr{G}] - g^*.$$

For each $t \in I$, let

$$\psi_1(s, t) = \tilde{g}_x(z(t), t) Y(s, t) \quad \text{for all } s \in [t_1, t], \tag{4.75}$$

and let

$$\psi_2(s) = \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} Y(s, t_2) \quad \text{for all } s \in I. \tag{4.76}$$

Define ψ_3 as follows:

$$\psi_3(s) = \psi_2(s) + \int_{s}^{t_2} \psi_1(s, t)\, d\lambda(t) \quad \text{for all } s \in I. \tag{4.77}$$

Then, (4.74) can be rewritten in the following form (after interchanging the order of integration in the first term):

$$\int_{t_1}^{t_2} \psi_3(s) \left\{ \delta g(z(s), s, s) + \int_{t_1}^{s} \delta g_s(z(\zeta), \zeta, s)\, d\zeta \right\} ds \leqq 0 \quad \text{for all } \delta g \in [\mathscr{G}] - g^*. \tag{4.78}$$

If we integrate by parts the second term in the right-hand side of (4.77), and use (4.75), (4.71) and (4.3), then we obtain

$$\psi_3(s) = \psi_2(s) - \lambda(s)\tilde{g}_x(z(s), s) - \int_s^{t_2} \lambda(t)\frac{\partial \psi_1(s,t)}{\partial t}\,dt \quad \text{for all } s \in I. \tag{4.79}$$

If we put

$$\bar{\psi}(s) = \psi_3(s) + \lambda(s)\tilde{g}_x(z(s), s) \quad \text{for all } s \in I, \tag{4.80}$$

then it can be easily seen that condition (f) of Theorem 2 is an immediate consequence of (4.78), (4.80) and the fact that $\delta g = g - g^*$.

We shall now prove condition (c) of Theorem 2. It is obvious that (see (4.80) and (4.79))

$$\bar{\psi}(s) = \psi_2(s) - \int_s^{t_2} \lambda(t)\frac{\partial \psi_1(s,t)}{\partial t}\,dt. \tag{4.81}$$

But, by the definitions of ψ_2 and ψ_1, both ψ_2 and $\psi_1(\cdot, t)$ are absolutely continuous on I. Hence, $\bar{\psi}(s)$ is absolutely continuous on I. Let us take the derivative of $\bar{\psi}(s)$ with respect to s. We have that

$$\frac{d\bar{\psi}(s)}{ds} = \frac{d\psi_2(s)}{ds} + \lambda(t)\frac{\partial \psi_1(s,t)}{\partial t}\bigg|_{t=s} - \int_s^{t_2} \lambda(t)\frac{\partial}{\partial s}\frac{\partial \psi_1(s,t)}{\partial t}\,dt \tag{4.82}$$

for almost all $s \in I$.

It follows from (4.76), (4.9), (4.75) and (4.81) that

$$\begin{aligned}\frac{d\bar{\psi}(s)}{ds} = {} & -\bar{\psi}(s)g_x^*(z(s), s, s) - \int_s^{t_2} [\bar{\psi}(\zeta) - \lambda(\zeta)\tilde{g}_x(z(\zeta), \zeta)]g_{x\zeta}^*(z(s), s, \zeta)\,d\zeta \\ & + \lambda(s)[\tilde{g}_{xx}(z(s), s)\dot{z}(s) + \tilde{g}_{xs}(z(s), s) + \tilde{g}_x(z(s), s)g_x^*(z(s), s, s)]\end{aligned} \tag{4.83}$$

for almost all $s \in I$.

This is exactly the same differential integral equation as in condition (c) of Theorem 2.

Now, we want to prove conditions (d) and (e). Take $\delta x(\cdot\,; \xi, \xi, 0) \in \mathcal{K}$. Then (4.73), (4.11) and (4.12) imply that

$$\left\{\int_{t_1}^{t_2} \tilde{g}_x(z(t), t)Y(t_1, t)\,d\lambda(t) + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} Y(t_1, t_2)\right\}\xi \leqq 0$$

for all $\xi \in R^n$.

This is possible only if

$$\int_{t_1}^{t_2} \tilde{g}_x(z(t), t)Y(t_1, t)\,d\lambda(t) + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} Y(t_1, t_2) = 0. \tag{4.84}$$

It then follows from the above equality, (4.75) and (4.76) that

$$\int_{t_1}^{t_2} \psi_1(t_1, t)\,d\lambda(t) + \psi_2(t_1) = -\sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1}.$$

By virtue of (4.71), (4.75) and (4.81), we conclude immediately that

$$\bar{\psi}(t_1) = -\sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} + \lambda(t_1)\tilde{g}_x(z(t_1), t_1).$$

Since $\lambda(t_2) = 0$, it is a consequence of (4.81), (4.77) and (4.76) that (also see (4.3))

$$\bar{\psi}(t_2) = \psi_3(t_2) = \psi_2(t_2) = \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2}.$$

If we choose $\delta x(\cdot\,; 0, \delta h, 0) \in \mathscr{K}$, then it follows from (4.73), (4.11) and (4.12) that

$$\int_{t_1}^{t_2} \tilde{g}_x(z(t), t) \int_{t_1}^{t} Y(s, t)\, \delta h(s)\, ds\, d\lambda(t) + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} \int_{t_1}^{t_2} Y(s, t)\, \delta \dot{h}(s)\, ds \leqq 0 \tag{4.85}$$

$$\text{for all } \delta h \in \mathscr{H} - h^* \quad \text{such that } \delta h(t_1) = 0.$$

Thus, condition (c) is a consequence of (4.85), (4.75)–(4.77) and (4.81).

Since conditions (a) and (b) are the parts of our early statements, the only thing left to be proved is condition (g).

For each $x \in C$, let us define a function W_x on I as follows:

$$W_x(t) = x(t) - \int_{t_1}^{t} g_x^*(z(s), s, t)x(s)\, ds \quad \text{for all } t \in I. \tag{4.86}$$

It is clear that the $\dot{W}_x$ cannot be all equal to zero for all $x \in C$. Based on Lemma 4.1 (where $\delta g = 0$) and (4.86), we can put x in the following form:

$$x(t) = Y(t_1, t)x(t_1) + \int_{t_1}^{t} Y(s, t)\dot{W}_x(s)\, ds \quad \text{for all } t \in I. \tag{4.87}$$

Let us define $\tilde{l}$ by

$$\tilde{l} = l_{-\mu-1} + \sum_{i=-\mu}^{m} a^i l_i.$$

By virtue of (4.70), $\tilde{l} \neq 0$. Hence $\tilde{l}(x) \neq 0$ for some $x \in C$. But,

$$\tilde{l}(x) = \int_{t_1}^{t_2} \tilde{g}_x(z(t), t)x(t)\, d\lambda(t) + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} x(t_1) + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} x(t_2),$$

or,

$$\begin{aligned}
\tilde{l}(x) &= \int_{t_1}^{t_2} \tilde{g}_x(z(t), t)\left\{Y(t_1, t)x(t_1) + \int_{t_1}^{t} Y(s, t)\dot{W}_x(s)\, ds\right\} d\lambda(t) + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} x(t_1) \\
&\quad + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2}\left\{Y(t_1, t_2)x(t_1) + \int_{t_1}^{t_2} Y(s, t_2)\dot{W}_x(s)\, ds\right\} \\
&= \left\{\int_{t_1}^{t_2} \tilde{g}_x(z(t), t)Y(t_1, t)\, d\lambda(t) + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_1} + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} Y(t_1, t_2)\right\} x(t_1) \\
&\quad + \int_{t_1}^{t_2} \int_{s}^{t_2} \tilde{g}_x(z(t), t)Y(s, t)\dot{W}_x(s)\, d\lambda(t)\, ds + \sum_{i=-\mu}^{m} \alpha^i \frac{\partial \chi_i}{\partial \xi_2} \int_{t_1}^{t_2} Y(s, t)\dot{W}_x(s)\, ds.
\end{aligned}$$

The above equality, together with (4.75)–(4.77) and (4.84), implies that

$$\tilde{l}(x) = \int_{t_1}^{t_2} \psi_3(s)\dot{W}_x(s)\,ds \quad \text{for all } x \in C.$$

Since $\dot{W}_x$ is not equal to zero for all $x \in C$ and $\tilde{l} \neq 0$, it implies that ψ_3 is not equal to zero on some subset of I of positive measure; i.e., by virtue of (4.80), $\bar{\psi}(s) - \lambda(s)\tilde{g}_x(z(s), s)$ is not equal to zero in some subset of I of positive measure.

This completes the proof of Theorem 2.

5. Acknowledgment. The author would like to thank Dr. L. W. Neustadt and Dr. L. S. Yeh for their many comments on the original version of this paper.

REFERENCES

[1] L. S. PONTRYAGIN, V. G. BOLTYANSKI, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[2] R. V. GAMKRELIDZE, *On some external problems in the theory of differential equations with applications to the theory of optimal control*, this Journal, 3 (1965), pp. 106–128.

[3] L. 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.

[4] ———, *An abstract variational theory with applications to a broad class of optimization problems II: Application*, this Journal, 5 (1967), pp. 90–137.

[5] H. HALKIN AND L. W. NEUSTADT, *General necessary conditions for optimization problems*, Proc. Nat. Acad. Sci., 56 (1966), pp. 1066–1071.

[6] A. FRIEDMAN, *Optimal control for hereditary processes*, Arch. Rational Mech. Anal., 15 (1964), pp. 396–416.

[7] A. HALANAY, *Optimal controls for systems with time lag*, this Journal, 6 (1967), pp. 215–234.

[8] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[9] F. G. TRICOMI, *Integral Equations*, Interscience, New York, 1965.

[10] S. C. HUANG, *Optimal control problems with retardations and restricted phase coordinates*, J. Optimization Theory and Applications, 3 (1969), pp. 316–360.

[11] V. R. VINOKUROV, *Optimal control of processes described by integral equations. I*, this Journal, 7 (1969), pp. 324–326.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

ON PURSUIT WITH CURVATURE CONSTRAINTS*

G. T. RUBLEIN†

Abstract. A theorem is proved extending results of Cockayne on pursuit with curvature constraints. Let two points (pursuer and evader) move in Euclidean 3-space with constant speeds. Provided the pursuer has greater speed and greater normal acceleration, it is shown that pursuit is always successful. The methods used are similar to Cockayne's. The pursuer, by some preliminary maneuvers, sets up a condition where he is leaving the line of sight in the same direction and with the same speed as the evader. It is shown that from this instant, the pursuer can, without violating constraints, keep the line of sight parallel to the original and ultimately collide with the evader.

1. Introduction. This note extends to three dimensions some results of Cockayne [1] giving sufficient conditions for capture of a slow but maneuverable evader by a speedy but clumsy pursuer.

Following the notation of [1], we have two points P (pursuer) and E (evader) moving in Euclidean 3-space with constant speeds w_1 and w_2, respectively. We require that their paths be continuously differentiable curves with piecewise continuous curvatures k_1 for P and k_2 for E, satisfying $|k_1| \leqq 1/R_1$ and $|k_2| \leqq 1/R_2$, where R_1 and R_2 are constants. A *state* of P and E at time t is the set of positions and velocities held by P and E at time t. *Capture* of E by P is defined to be coincidence of positions of P and E.

We prove the following theorem.

THEOREM. *P can capture E from any initial state if*

(A) $$w_1 > w_2, \quad \textit{and}$$

(B) $$w_1^2/R_1 > w_2^2/R_2.$$

Cockayne gives maneuvers for E's escape in case (A) fails or in case $w_1^2/R_1 < w_2^2/R_2$. If both P and E are constrained to move in the plane, then he shows that a weak inequality can be used in (B). The author does not know whether an equality in (B) yields capture in 3-space.

As in [1], we assume that in P's efforts to capture E, he has complete information concerning E's instantaneous motion (i.e., second derivative) in addition to his knowledge of the state of P and E.

2. Proof. Our procedure is to give practical maneuvers for the capture. We make regular use of a moving right-handed coordinate system in which the z-axis points along the line of sight from P to E and the x, z-plane contains $\mathbf{v}_1$, the velocity vector of P, $\mathbf{v}_1$ having a nonnegative x-component. If $\mathbf{v}_2$ is the velocity vector of E and $\mathbf{a}_1$ and $\mathbf{a}_2$ are the acceleration vectors of P and E respectively, we write in the moving frame $\mathbf{v}_1 = (\dot{x}_1, 0, \dot{z}_1)$, $\mathbf{v}_2 = (\dot{x}_2, \dot{y}_2, \dot{z}_2)$, $\mathbf{a}_1 = (\xi_1, \eta_1, \zeta_1)$, $\mathbf{a}_2 = (\xi_2, \eta_2, \zeta_2)$. $\mathbf{L}$ is the vector from P to E so that $\mathbf{L} = |\mathbf{L}|\mathbf{k}$, where $\mathbf{k}$ is the unit vector along the positive z-axis, and $d\mathbf{L}/dt = \mathbf{v}_2 - \mathbf{v}_1$.

The constant speeds along both paths give

(1) $$\mathbf{v}_1 \cdot \mathbf{a}_1 = \mathbf{v}_2 \cdot \mathbf{a}_2 = 0.$$

* Received by the editors October 29, 1970.

† Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23185. This work was supported in part by the National Aeronautics and Space Administration under contract NAS1-9461-6.

The curvature constraint on P gives $\xi_1^2 + \eta_1^2 + \zeta_1^2 \leqq w_1^4/R_1^2$, and using (1) we have

$$\xi_1^2 \leqq (w_1^4/R_1^2 - \eta_1^2)\dot{z}_1^2/w_1^2. \tag{2}$$

LEMMA 1. *If condition* (A) *holds, then, from any initial state, P can move aginst all opposition so that the distance* $|\mathbf{L}|$ *becomes arbitrarily large.*

Proof. Obvious.

LEMMA 2. *If condition* (A) *holds, then from any initial state, P can force the motion into a state satisfying*:

(i) $|\mathbf{L}|$ *is arbitrarily large,*

(ii) $\mathbf{v}_1$ *points along* $\mathbf{L}$.

Proof. Let λ be the angle made by $\mathbf{v}_1$ and $\mathbf{L}$. Then $w_1 \cos\lambda = \mathbf{v}_1 \cdot \mathbf{L}/|\mathbf{L}|$. A computation gives

$$w_1 \sin\lambda \, d\lambda/dt = -[(\mathbf{a}_1 \cdot \mathbf{L}) - w_1^2 \sin^2\lambda + \mathbf{v}_2 \cdot (\mathbf{v}_1 - w_1 \cos\lambda \mathbf{k})]/|\mathbf{L}|. \tag{3}$$

One verifies directly that $|\mathbf{v}_2 \cdot (\mathbf{v}_1 - w_1 \cos\lambda \mathbf{k})| \leqq w_1 w_2 \sin\lambda$, and that $|\mathbf{k} - \mathbf{v}_1 \cos\lambda/w_1| = \sin\lambda$. Hence, P may select $\mathbf{a}_1 = [\mathbf{k} - \mathbf{v}_1 \cos\lambda/w_1]w_1^2/(R_1 \sin\lambda)$. Then (1) is satisfied, and, for large $|\mathbf{L}|$, $w_1 \, d\lambda/dt$ is dominated by $-\mathbf{a}_1 \cdot \mathbf{k}/\sin\lambda = -w_1^2/R_1$. The indicated maneuver for P is therefore to retreat a large distance from E and then take $\mathbf{a}_1$ as above. In time no larger than $2w_1/(\pi R_1)$ (for sufficiently large $|\mathbf{L}|$), λ will become 0. Note that when $\lambda = 0$ is attained, we may still assume that $|\mathbf{L}|$ is arbitrarily large.

LEMMA 3. *If* (A) *and* (B) *hold, then from any initial state, P can force the motion into a state satisfying*:

(iii) *the component of* $\mathbf{v}_1$ *along* $\mathbf{L}$ *is positive,*

(iv) $\mathbf{v}_1, \mathbf{v}_2, \mathbf{L}$ *are coplanar,*

(v) *the components of* $\mathbf{v}_1$ *and* $\mathbf{v}_2$ *perpendicular to* $\mathbf{L}$ *are equal.*

Proof. Making use of Lemma 2, we may assume that initially $|\mathbf{L}|$ is large and that $\mathbf{v}_1$ points along $\mathbf{L}$. If $\mathbf{v}_2$ also points along $\mathbf{L}$ there is nothing to do, so we may confine ourselves to the case where $\dot{x}_2 > 0$ (the x, z-plane may be chosen to contain $\mathbf{v}_2$). Now, suppose that at any subsequent instant $|\mathbf{L}|$ is large, that (iii) and (iv) are satisfied and further that

$$\dot{x}_1 \leqq \dot{x}_2 \tag{v'}$$

holds. By (iv), the two members of (v') are the respective components of $\mathbf{v}_1$ and $\mathbf{v}_2$ perpendicular to $\mathbf{L}$. Our objective here is to describe a maneuver of P which maintains (iii) and (iv) and accomplishes (v).

Let $\mathbf{m} = (\mathbf{v}_1 \times \mathbf{v}_2) \cdot \mathbf{L}$. At the given instant, $\mathbf{m} = \mathbf{0}$. Since $d\mathbf{m}/dt = \mathbf{L} \cdot (\mathbf{a}_1 \times \mathbf{v}_2 + \mathbf{v}_1 \times \mathbf{a}_2) = |\mathbf{L}|(-\eta_1 \dot{x}_2 + \dot{x}_1 \eta_2)$ at this instant, P can maintain (iv) (i.e., can maintain $\mathbf{m} = \mathbf{0}$) by setting

$$\eta_1 = \eta_2 \dot{x}_1/\dot{x}_2. \tag{4}$$

By (v'), $|\eta_1| \leqq |\eta_2|$.

Next, let h_1 and h_2 be the components of $\mathbf{v}_1$ and $\mathbf{v}_2$ perpendicular to $\mathbf{L}$. (In our coordinate system, these are instantaneously $\dot{x}_1$ and $\dot{x}_2$). If we let p_1 and p_2 be the respective components of $\mathbf{v}_1$ and $\mathbf{v}_2$ along $\mathbf{L}$, then we have $h_i^2 = w_i^2 - p_i^2$,

$i = 1, 2$. We can therefore compute dh_1/dt from

$$\begin{aligned} dh_1/dt &= -(1/(2h_1))\, dp_1^2/dt \\ &= -(1/h_1)(\mathbf{v}_1 \cdot \mathbf{k}/|\mathbf{L}|)(\mathbf{a}_1 \cdot \mathbf{L} + \mathbf{v}_1 \cdot \mathbf{v}_2 - w_1^2 + (\mathbf{v}_1 \cdot \mathbf{k})^2 - (\mathbf{v}_1 \cdot \mathbf{k})(\mathbf{v}_2 \cdot \mathbf{k})). \end{aligned} \tag{5}$$

Examination of (5) shows that for large $|\mathbf{L}|$, dh_1/dt is dominated by $-(1/h_1) \cdot (\mathbf{v}_1 \cdot \mathbf{k})(\mathbf{a}_1 \cdot \mathbf{k})$, provided this term is bounded away from 0. In the moving frame, this term is $-(1/\dot{x}_1)\dot{z}_1\zeta_1 = \xi_1$. Using a similar expression for dh_2/dt, we finally get that for large $|\mathbf{L}|$, $d(h_2 - h_1)/dt$ is dominated by $\xi_2 - \xi_1$ (and is negative) provided this term is bounded below 0.

Let P turn with maximum curvature using (4) with ξ_1 positive. Then, from (2), we get

$$\begin{aligned} \xi_1 &= \sqrt{(w_1^4/R_1^2 - \eta_1^2)}|\dot{z}_1|/w_1 > \sqrt{(w_2^2/R_2^4 - \eta_2^2)}|\dot{z}_1|/w_1 \\ &\geqq |\xi_2 \dot{z}_1/\dot{z}_2| w_2/w_1 . \end{aligned} \tag{6}$$

Now it is easy to verify that when (A) and (v′) hold, $|\dot{z}_1/\dot{z}_2| w_2/w_1 > 1$. Hence, the last quantity in (6) is at least as large as $|\xi_2|$. Finally, we observe that the first inequality in (6) is a bounded inequality by virtue of (B). Hence, ξ_1 is bounded above $|\xi_2|$ so that $d(h_2 - h_1)/dt$ is (for large $|\mathbf{L}|$) a quantity bounded below 0.

It follows that in time no larger than some constant C, $h_2 - h_1$ will become 0. From (A), one sees immediately that at the first instant when this occurs $\dot{z}_1 > 0$.

To complete the proof of the theorem, we make use of Lemma 3 to assume that initially $\mathbf{L}$, $\mathbf{v}_1$ and $\mathbf{v}_2$ are coplanar, that $\dot{z}_1 > 0$, and that $\dot{x}_1 = \dot{x}_2$. It follows from (A) that $\dot{z}_1 > |\dot{z}_2|$. From this point, we take

$$\xi_1 = \xi_2, \qquad \eta_1 = \eta_2. \tag{7}$$

From (1), $|\zeta_1| = |\dot{x}_1\xi_1/\dot{z}_1| < |\dot{x}_2\xi_2/\dot{z}_2| = |\zeta_2|$. The curvature constraint is therefore satisfied by virtue of

$$\xi_1^2 + \eta_1^2 + \zeta_1^2 < \xi_2^2 + \eta_2^2 + \zeta_2^2 \leqq w_2^4/R_2^2 < w_1^4/R_1^2.$$

Equation (7) implies that (iii), (iv) and (v) are maintained. Condition (A) implies that $d|\mathbf{L}|/dt = \dot{z}_2 - \dot{z}_1$ is bounded below 0 so that collision between P and E is insured.

3. Remark. It is pointed out in [1] that for plane pursuit the usual information patterns of a differential game are sufficient to insure collision. It appears to the author that this is not the situation for three-dimensional pursuit. Although it will still not bring us formally into the realm of a differential game, we might get around the rather impractical looking information pattern used here as follows: Define *collision* to be a state where $|\mathbf{L}| < \delta$, for some $\delta > 0$. Now we instruct P to observe only the state of E (and himself) at time t. Then by differentiating, he can deduce E's acceleration at time $t - \delta/(w_1 + w_2)$. Using the above prescription, P can collide (in the earlier sense) with a "phantom" who trails E with a time lag $\delta/(w_1 + w_2)$. He will then collide (in the new sense) with E.

REFERENCES

[1] E. Cockayne, *Plane pursuit with curvature constraints*, SIAM J. Appl. Math., 15 (1967), pp. 1511–1516.

[2] R. Isaacs, *Differential Games*, John Wiley, New York, 1965.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

GAMES WITH A "LIFE-LINE". THE CASE OF l-CAPTURE*

YU. G. DUTKEVICH AND L. A. PETROSYAN†

Abstract. We consider an antagonistic pursuit game with a life-line with simple motions taking place in a given convex closed set in the plane. The evader E is considered to have been captured if the Euclidean distance $\rho(P, E)$, where P is the pursuer, does not exceed l, where l is some preassigned positive number. Optimal strategies which make up an equilibrium point are found for both players.

1. Problem statement. A family of antagonistic games is considered. Each of these games is a model of a pursuit in some closed convex set $S \subset R^2$. Two points, a pursuer P and an evader E, which have constant (linear) velocities u and v (with $u > v$), are moving in S. They are able to change the direction of their motion at each instant of time. The evader E is considered to have been captured as soon as his distance from P becomes $\leqq l$ (here, $l > 0$). The number l is called the capture radius. The aim of E is to reach the boundary of S before an l-capture by the pursuer. At each instant of time, P has information about his position, the position of E, and the direction of the velocity of E at that time. In the model which we shall consider, we shall assume that the trajectories of the players are piecewise-continuously differentiable. The piecewise-continuous differentiability of the players' trajectories guarantees that the directions of the players' velocities exist at all points of the trajectory, with the possible exception of a finite number of points at which there exists a right-hand derivative (i.e., a semi-tangent) about which we shall assume that the player P obtains information at each instant of time. The evader E has information about his and P's positions. The "point-wise" capture case ($l = 0$) was solved in [1].

The kinematic equations have the form

$$\dot{x}_i = \varphi_i,$$

$$\dot{y}_i = \psi_i, \qquad i = 1, 2,$$

$$\varphi_1^2 + \varphi_2^2 = u^2, \qquad \psi_1^2 + \psi_2^2 = v^2.$$

Here, $x = (x_1, x_2)$ is the position of P, $y = (y_1, y_2)$ is the position of E, $\varphi = (\varphi_1, \varphi_2)$ is the control variable of P, and $\psi = (\psi_1, \psi_2)$ is the control variable of E.

As is the usual convention in the theory of differential games, a function $\varphi(x, y, \psi)$, which satisfies the condition $\varphi_1^2 + \varphi_2^2 = u^2$ and which assigns to the current information of P some choice of the control variable, is said to be a strategy for the player P. Similarly, we shall call a function $\psi(x, y)$, which satisfies the condition $\psi_1^2 + \psi_2^2 = v^2$, a strategy for the player E. We denote by Φ and Ψ the classes $\Phi = \{\varphi\}$ and $\Psi = \{\psi\}$ of all possible functions that satisfy the following conditions:

* Originally published in *Vestnik Leningradskogo Universiteta*, (1969), no. 13, pp. 31–38. Submitted April 11, 1968. This translation into English has been prepared by K. Makowski.

Translated and printed for this Journal under a grant-in-aid from the National Science Foundation.

† Mathematics and Mechanics Department, Leningrad State University, Leningrad, U.S.S.R.

(i) For any initial conditions $x_0, y_0 \in S$, and for any pair of functions $\varphi \in \Phi$, $\psi \in \Psi$, the system of ordinary differential equations

$$\begin{aligned} \dot{x}_i &= \varphi_i(x, y, \varphi(x, y, \psi)), \\ \dot{y}_i &= \psi_i(x, y, \psi(x, y)), \qquad i = 1, 2, \end{aligned} \tag{1.1}$$

has a unique solution $x(t)$, $y(t)$.

(ii) Let $x(t)$ be the solution of the system of differential equations (1.1) in a situation (φ, ψ), from initial positions $x_0, y_0 \in S$, and let

$$t_{S_P} = \inf\{t : x(t) \notin S\},$$
$$t_{S_E} = \inf\{t : y(t) \notin S\}.$$

Then, for all

$$t > t_{S_P}, \quad x(t) \notin S,$$
$$t > t_{S_E}, \quad y(t) \notin S.$$

Further, let

$$t_P = \min\{t : \rho(x(t), y(t)) = l\},$$

where $\rho(x, y)$ denotes Euclidean distance (if no such t exists, we set $t_P = \infty$).

The payoff function. Let $x(t)$, $y(t)$ be the trajectories of the players P and E which originate at the initial positions $x_0, y_0 \in S$ in a situation (φ, ψ). Then the payoff function (the payoff for P) is

$$K(x_0, y_0; \varphi, \psi) = \begin{cases} +1 & \text{if } t_P \leqq t_{S_E}, \quad t_{S_E} \neq \infty, \\ 0 & \text{if } t_P = t_{S_E} = \infty, \\ -1 & \text{if } t_P > t_{S_E}. \end{cases}$$

Having defined the sets of the players' strategies and the payoff function, we assign a certain family of games in normal form. The games depend on the initial positions $x_0, y_0 \in S$. Each game from this family will be denoted by $\Gamma(x_0, y_0)$.

It follows from the form of the payoff function that $\Gamma(x_0, y_0)$ is a game of kind (for the definition of a game of kind, see [2]) in which P strives to closely approach P before the latter crosses the boundary of S. We shall assume that the game is antagonistic, i.e., that the payoffs for E and for P are the negatives of one another (for antagonistic games in more detail, see [3]).

The assumption concerning the discrimination against E is a natural one in order to assure the existence of the value of the game (see, e.g., L.S. Pontryagin [4]).

Let $\bar{\varphi}$ be a fixed strategy for P which has the property that, in any situation $(\bar{\varphi}, \psi)$, an l-capture of E in the game $\Gamma(x_0, y_0)$ can be realized in the entire space. Let $C_{\bar{\varphi}}(x_0, y_0)$ be the set of all positions of E at the instant of l-capture in a situation $(\bar{\varphi}, \psi)$ (the position of E at the instant of l-capture will in the sequel also be called a "meeting" point). Obviously, if $C_{\bar{\varphi}}$ has a nonempty intersection with the complement of S, then P, using the strategy $\bar{\varphi}$, cannot guarantee an l-capture of E in S. Therefore, the player E in this case can always choose a strategy ψ^*

such that an l-capture in the situation $(\bar{\varphi}, \psi^*)$ takes place in the complement of S. There may exist many such strategies.

THEOREM 1. *Let the following conditions hold*:

(i) *The intersection of* $C_{\bar{\varphi}}(x_0, y_0)$ *with the complement of* S *is not empty.*

(ii) *In the class of strategies* ψ *such that for initial position* (x_0, y_0) *the situation* $(\bar{\varphi}, \psi)$ *results in* l*-capture with "meeting" point not in* S*, let there exist a strategy* ψ^* *such that in the situation* $(\bar{\varphi}, \psi^*)$*, the player* P *realizes* l*-capture in minimum (finite) time.*

Then there exists a strategy $\tilde{\psi}^*$ *such that the situation* $(\bar{\varphi}, \tilde{\psi}^*)$ *is a saddle-point in the game* $\Gamma(x_0, y_0)$ *and the value of the game equals* -1.

(This means that if the conditions of the theorem hold and E uses $\tilde{\psi}^*$, the l-capture of E in S is impossible regardless of the pursuer's strategy.)

Proof. Let $x^*(t)$, $y^*(t)$ be the trajectories of P and E, respectively, in the situation $(\bar{\varphi}, \psi^*)$ of the game $\Gamma(x_0, y_0)$. Let us denote by $\tilde{\psi}^*$ the strategy for E under which he, no matter what P does, chooses, at each instant of time, the direction of his motion along the trajectory $y^*(t)$. No matter what φ is, an l-capture in the situation $(\varphi, \tilde{\psi}^*)$ cannot take place before the time

$$t_P(\bar{\varphi}, \psi^*) = t_P = t_P(\bar{\varphi}, \tilde{\psi}^*),$$

since, in the situations $(\bar{\varphi}, \psi^*)$ or $(\bar{\varphi}, \tilde{\psi}^*)$, P realizes the fastest movement to the "meeting" point. This means that

$$t_P(\varphi, \tilde{\psi}^*) \geqq t_P(\bar{\varphi}, \tilde{\psi}^*) > t_{S_E}$$

for all strategies φ of P. The theorem has been proved.

When $C_{\bar{\varphi}}(x_0, y_0)$ is contained in S, an l-capture of E in S is always possible, if P uses the strategy $\bar{\varphi}$. In this case, $\bar{\varphi}$ turns out to be an optimal strategy for P (it may not be unique), the value of the game is 1 and any strategy is optimal for E.

Thus, to solve the game, it is sufficient to establish either the existence of a strategy ψ^* for E which satisfies the conditions of Theorem 1, or the existence of a strategy φ^* such that $C_{\varphi^*}(x_0, y_0)$ is contained in S.

Let us now pass to constructing the pair of strategies (φ^*, ψ^*) which have the property just described.

Let E choose, from an initial position y_0, some constant control $\psi = \text{const}$. (i.e., let E move in a straight line uniformly along a ray $y_0 A$). For every such motion, P obviously has a unique constant control $\bar{\varphi}$ which guarantees him an l-capture of E from the position x_0 in minimum time. This control directs him to move along the ray xB aimed at the "meeting" point (see Fig. 1).

DEFINITION. The strategy φ^{Π} which, to every pair of positions x, y, and to each control ψ of E, corresponds the control $\bar{\varphi}$ which guarantees an l-capture of E from x in minimum time, provided that E adheres to a constant control ψ during the entire game, will be called the Π-*strategy* for P.

It is easy to see that, in the case $l = 0$, the strategy φ^{Π} coincides with the parallel pursuit strategy (see [1]).

We shall now determine the structure of $C_{\varphi\Pi}(x_0, y_0)$. Consider the subset $\bar{C}_{\varphi\Pi}(x_0, y_0)$ which is obtained from $C_{\varphi\Pi}(x_0, y_0)$ under the condition that E use only constant strategies ψ (i.e., E moves only along straight half-lines originating

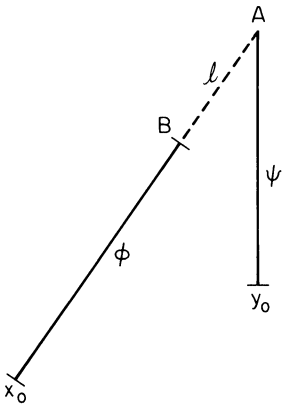


FIG. 1

at (y_0). We shall show that $C_{\varphi\Pi}(x_0, y_0)$ is a convex set and that $\bar{C}_{\varphi\Pi}(x_0, y_0)$ is its boundary. The proof will be carried out in § 2 (see Theorem 4).

THEOREM 2. *Assume that $C_{\varphi\Pi}(x_0, y_0)$ intersects the complement of S. Then, for a* Π*-strategy to satisfy the conditions of Theorem* 1, *it is sufficient that there exist a strategy ψ^* of E such that E moves along a half-line and such that the "meeting" point in the situation (φ^Π, ψ^*) belongs to the complement of S.*

Proof. The theorem follows at once from the fact that, in the situation (φ^Π, ψ^*) (see the definition of a Π-strategy), P will also move along the half-line which passes through x_0 and the "meeting" point, i.e., he will realize the fastest motion to the "meeting" point.

Obviously, if the condition $C_{\varphi\Pi} \cap S' \neq \Lambda$ is satisfied, then Theorem 4 of § 2 guarantees the existence of such a strategy (such a half-line passing through y_0 and the "meeting" point). We obtain the following theorem.

THEOREM 3. *To avoid an l-capture of E in S, it is necessary and sufficient that $C_{\varphi\Pi}(x_0, y_0)$ have a nonempty intersection with the complement of S.*

A construction of $C_{\varphi\Pi}(x_0, y_0)$ is given in § 2.

2. A basic lemma. Let us introduce Cartesian coordinates x and y on a plane. At $t = 0$, let the pursuer have coordinates $(0, 0)$ and the evader E have coordinates $(a, 0)$, where $a > l$. For ease of notation, we shall assume that $u = 1$ and $v = \alpha$, where $\alpha < 1$. At the time t, let P be at the point $P(t) = (x_P^t, y_P^t)$ and E at the point $E(t) = (x_E^t, y_E^t)$.

Let P and E move on straight lines along rays L_P and L_E, and let them "meet" at a time t at the point w $(w = L_P \cap L_E)$. Then the coordinates (x, y) of w satisfy the relations

$$\begin{aligned} x^2 + y^2 &= (t + l)^2, \\ (x - a)^2 + y^2 &= (\alpha t)^2, \\ t &> 0. \end{aligned} \tag{2.1}$$

Equations (2.1) define a curve in the x, y-plane which is called the oval of Descartes [5]. Since $t > 0$, we obtain a part of the oval of Descartes. Let us denote this part by D.

Thus, moving on straight lines, P and E "meet" on the curve D.

At a time τ, let P and E change the directions of their motions (P changes its direction so as to realize an l-capture). Then there emerges a new "meeting" curve $D(\tau)$ which is given by the equations

$$
\begin{aligned}
&(x - x_P^\tau)^2 + (y - y_P^\tau)^2 = (t - \tau + l)^2, \\
&(x - x_E^\tau)^2 + (y - y_E^\tau)^2 = [\alpha(t - \tau)]^2, \\
&\tau < \tilde{t}
\end{aligned}
\tag{2.2}
$$

($\tilde{t}$ is the "meeting" time of P and E if they move along L_P and L_E, respectively). Let us note that, for $\tau = 0$, (2.2) become (2.1).

The following basic lemma holds.

LEMMA. *$D(\tau)$ is a closed, bounded, convex curve for $\alpha < 1$, $a > l$, and $\tau < \tilde{t}$. Moreover, for $\tau > 0$, $D(\tau)$ lies in the closed domain bounded by the curve $D = D(0)$ (see Fig. 2).*

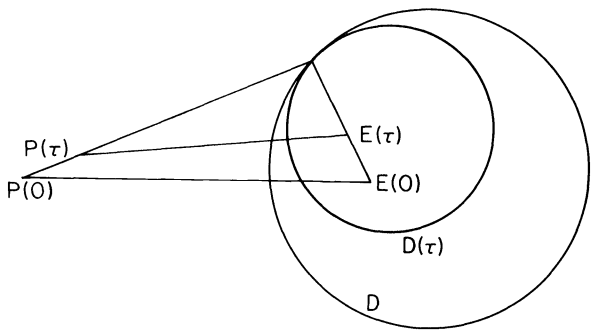


FIG. 2

Before proving the lemma, let us present, without proof, three simple remarks.

Remark 1. If any straight line intersects a curve K at no more than two points, then K is strictly convex.

Remark 2. Let there be given upward directed branches of two hyperbolas as follows:

$$
y^2 - x^2 = m^2,
$$

$$
(y - \beta)^2 - \mu^2(x - \alpha)^2 = n^2, \quad \mu > 1, \quad \beta \geqq 0, \quad |\alpha| \geqq \beta.
$$

These branches have at most two points in common.

Remark 3. Under a homothetic transformation of one branch γ of a hyperbola, with the center of the homothety at a point $M \in \gamma$, γ is transformed into a branch γ' of a hyperbola, where γ and γ' are tangent at the point $M \in \gamma$. If the coefficient of the homothety $k < 1$, then the convex region G bounded by γ contains γ'. If $k > 1$, then γ' lies outside of G.

Proof of the lemma. In the Euclidean space with coordinates (x, y, t), (2.2) define right circular cones with axes parallel to the t-axis. Under the condition $\tau < t$, each of (2.2) defines the upper (i.e., directed to the side of increasing t) parts $K(P, \tau)$ and $K(E, \tau)$ of these cones, and the curve $D(\tau)$ is the projection onto the x, y-plane of the curve of intersection of these cones.

(i) Let us show that $D(\tau)$ is convex. To this effect (Remark 1), it is sufficient to verify that any straight line intersects D at no more than two points. Let p be a straight line in the x, y-plane, and let π be the plane passing through p and parallel to the t-axis. The sets $\pi \cap K(P, 0)$ and $\pi \cap K(E, 0)$ are the upper branches of two hyperbolas which, according to Remark 2, have at most two points in common. The convexity has been proved.

(ii) Let us prove that, for $\tau > 0$, $D(\tau)$ lies in the closed (convex) region bounded by $D = D(0)$. To this effect, it is sufficient to establish that, for any straight line passing through the point $w = D(\tau) \cap D$, the chord in $D(\tau)$ is a part of the chord sliced off D by the same line.

Let P and E move along the straight lines L_P and L_E, respectively. At a time t, let there correspond to them the points $\tilde{P}(t) \in K(P, 0)$ and $\tilde{E}(t) \in K(E, 0)$ whose projections are $P(t)$ and $E(t)$, respectively. The points $\tilde{P}(t)$ and $\tilde{E}(t)$ draw the rays $\tilde{L}_P \subset K(P, 0)$ and $\tilde{L}_E \subset K(E, 0)$ which intersect at $\tilde{w}$ (the projection of $\tilde{w}$ onto the x, y-plane is w). Since, at the time $t = \tau$, the directions of the motions of P and E change, there will analogously correspond to $P(t)$ and $E(t)$, for $t > \tau$, the points $\tilde{P}(t) \in K(P, \tau)$ and $\tilde{E}(t) \in K(E, \tau)$.

Let p be a straight line in a plane passing through w, and let π be a plane passing through p and parallel to the t-axis. Let us denote

$$\gamma(P, \tau) = \pi \cap K(P, \tau), \qquad \gamma(E, \tau) = \pi \cap K(E, \tau).$$

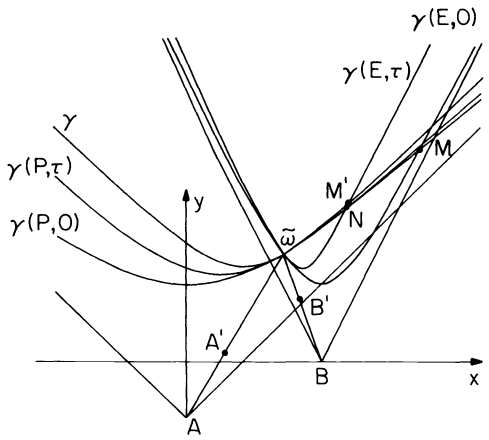


FIG. 3

The curves $\gamma(P, 0)$, $\gamma(E, 0)$, $\gamma(P, \tau)$ and $\gamma(E, \tau)$ are branches of hyperbolas lying in the plane π. All of them pass through $\tilde{w}$. The transverse axes of these hyperbolas are parallel to the t-axis. The center of the hyperbola $\gamma(P, 0)$ lies at the point A which is the orthogonal projection of $P(0)$ onto π. The center of the hyperbola $\gamma(P, \tau)$ lies at the point A' which is the orthogonal projection of $P(\tau)$ onto π. Therefore, A, A', and $\tilde{w}$ lie on a single straight line, which is the projection of the line $P(0)w$ onto π. Hence, it follows that $\gamma(P, \tau)$ is obtained from $\gamma(P, 0)$ by a homothety with coefficient $k_P = \tilde{w}A'/\tilde{w}A = (\tilde{t} - \tau + l)/(\tilde{t} + l)$, where $\tilde{t}$ is the z-coordinate of $\tilde{w}$. Similarly, let B and B' be the centers of the hyperbolas $\gamma(E, 0)$

and $\gamma(E,\tau)$. The points B, B', and $\tilde{w}$ lie on a single straight line. Therefore, $\gamma(E,\tau)$ is obtained from $\gamma(E,0)$ by a homothety with coefficient $k_E = \tilde{w}B'/\tilde{w}B = (\tilde{t}-\tau)/\tilde{t}$. It is clear that $k_E < k_P < 1$.

Let M and N be the second points of intersections of $\gamma(P,0)$ with $\gamma(E,0)$, and $\gamma(P,\tau)$ with $\gamma(E,\tau)$, respectively. We shall show that the projection of the line segment $\tilde{w}M$ onto the x,y-plane covers the projection of $\tilde{w}N$. Indeed, under the homothety with center at $\tilde{w}$ and coefficient k_E, $\gamma(E,0)$ is transformed into $\gamma(E,\tau)$, and $\gamma(P,0)$ is transformed into γ'. The second point of intersection M' of $\gamma(E,\tau)$ with γ' lies on the line segment $\tilde{w}M$. The curve $\gamma(P,\tau)$ is obtained from γ' by a homothety with coefficient $k = k_P/k_E > 1$, so that N (Remark 3) lies on the arc $\tilde{w}M'$ of the curve $\gamma(E,\tau)$. Since $\gamma(E,\tau)$ is projected onto the x,y-plane in a unique way, the projection of N lies on the projection of the line segment $\tilde{w}M$. The lemma has been proved.

Repeatedly applying the basic lemma, we can show that, in the situation (φ^{Π},ψ), where ψ is a strategy for E which directs him to move along a polygonal line, the "meeting" point belongs to the region bounded by the curve D (which we shall in the sequel call a D-oval). Passing to the limit we can further show that, in any situation (φ^{Π},ψ), the "meeting" point belongs to the region bounded by the curve D, i.e., the D-oval turns out to be the boundary of $C_{\varphi\Pi}(x_0,y_0)$. Let us formulate this result as a theorem.

THEOREM 4. *The D-oval which is the set of "meeting" points for straight-line movements of E coincides with the boundary of the set $C_{\varphi\Pi}(x_0,y_0)$.*

3. Some remarks and unsolved problems.

Remark 1. Let S be a convex set. Consider a "life-line" game in the complement of S. The set $C_{\varphi\Pi}(x,y)$ is then the same as the set of Theorem 4 of § 2. Therefore, all of the arguments remain in force, if, in a situation (φ^{Π},ψ), the trajectory of P does not cross the boundary of S before the game ends. In the contrary case, the Π-strategy will turn out to be inadmissible, since it will not satisfy conditions (i) and (ii) of § 1. This will not happen if $M(x,y)$, which is the convex hull of $C_{\varphi\Pi}(x,y) \cup \{x\}$, does not intersect S. In other words, the following theorem holds.

THEOREM 5. *Let $M(x,y)$ be contained in the complement of S. Then the Π-strategy is an optimal strategy for P, and an l-capture of E is always possible in the complement of S.*

Remark 2. In the case when, instead of one pursuer, a group of pursuers $P = \{P_1,\cdots,P_m\}$ take part in the game (acting as one player), and the minimum velocity u_i of P_i is greater than the velocity of E, the optimal strategy for P is the vector $\varphi^{\Pi} = \{\varphi^{\Pi_1},\cdots,\varphi^{\Pi_m}\}$, where φ^{Π_i} is the Π-strategy in the game with players P_i and E. A proof of this fact may be obtained by making use of the fact that the set of "meeting" points in a situation (φ^{Π},ψ), from initial positions $x_0 = (x_0^1,\cdots,x_0^m)$, y_0, is the intersection $\bigcap_{i=1}^m C_{\varphi\Pi}(x_0^i,y_0^i)$ of the corresponding sets in the game with players P_i and E.

Remark 3. Consider the game $\Gamma(x,y)$ from initial positions $x,y \in S$, such that $C_{\varphi\Pi}(x,y) \subset S$. Then, according to Theorem 2 of § 1, an l-capture of E is always possible in S. In this case, the following problem statement is a natural one. The player E strives to minimize the distance between himself and the boundary of S at the "meeting" point (he strives to be captured close to his "native

shore," the boundary of S). In this case, the saddle-point situation again has the form (φ^{Π}, ψ^*), where ψ^* is the strategy for E which directs him to move along the ray joining y_0 with the point of the oval of Descartes which is situated closest to the boundary of S.

Problems for future investigation:

(i) Solve the game with a "life-line" on an arbitrary smooth surface. Even the case $l = 0$ is interesting.

(ii) Solve the game with a "life-line" in a set which is not convex.

REFERENCES

[1] L. A. PETROSYAN, *Pursuit games with a "life-line,"* Vestnik Leningrad Univ., (1967), no. 13, pp. 76–86.

[2] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[3] J. C. C. MCKINSEY, *Introduction to the Theory of Games*, McGraw-Hill, New York, 1952.

[4] L. S. PONTRYAGIN, *On the theory of differential games*, Uspekhi Mat. Nauk, 21 (1966), pp. 219–274 (in Russian).

[5] A. A. SAVELOV, *Plane Curves*, Fizmatgiz, Moscow, 1960.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

ON DYNAMICAL SYSTEMS REALIZING STATIONARY WEIGHTING PATTERNS AND TIME-VARYING FEEDBACK SYSTEMS*

R. A. SKOOG†

Abstract. In this paper, an improved necessary and sufficient condition is obtained for a linear time-varying dynamical system to realize a stationary weighting pattern. These results are then used to show that certain time-varying feedback systems will have the input-output mapping of a time-invariant system only in trivial cases.

1. Introduction. In this paper, a necessary and sufficient condition is obtained in terms of the matrices A, B and C under which the linear time varying dynamic system (A, B, C) realizes a stationary weighting pattern (i.e., its input-output mapping is that of a time-invariant system). This result is a slight improvement over a result obtained by Silverman and Meadows [5]. These results are then used to show that certain linear time varying feedback systems cannot realize a time invariant input-output behavior except in the trivial cases when the time variations do not contribute to the input-output behavior.

The motivation for studying these problems lies primarily in the fact that when one is constrained to using only certain types of components in building a system, a considerably larger class of system functions can be realized when the component values are allowed to be time varying. Examples of this in network theory are given in [1]–[3]. Since synthesis of time-varying systems is most easily accomplished using the state variable formulation, the importance of conditions on A, B and C which guarantee that the weighting pattern will be stationary can be readily appreciated. With regards to feedback systems, the simplest form of a time-varying feedback system is one with a linear time invariant forward path and a time-varying gain in the feedback path. It would be most useful in stability theory and also in feedback system synthesis to be able to obtain a time-varying closed loop system which had the input-output behavior of a time-invariant system. The results here show, however, that this will only happen in a trivial way.

2. Preliminaries. The systems to be considered here are those having a representation in the form

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t), \end{aligned} \tag{2.1}$$

where the state $x(t)$ is a real n-vector, the input $u(t)$ is a real m-vector, and the output $y(t)$ is a real p-vector. The real matrices $A(t)$, $B(t)$ and $C(t)$ are respectively $n \times n$, $n \times m$ and $p \times n$. A system in the form of (2.1) will be denoted by (A, B, C).

* Received by the editors September 15, 1970, and in revised form April 21, 1971.

† Electronic Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts. Now at Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720. This work was supported in part by the National Science Foundation, under Grant GK-5786 and in part by the Lincoln Library of the Massachusetts Institute of Technology.

As is well known, the output y of system (2.1) is given by

$$y(t) = C(t)\Phi(t, t_0)x_0 + \int_{t_0}^{t} C(t)\Phi(t, \tau)B(\tau)u(\tau), \tag{2.2}$$

where x_0 is the initial state at time t_0 and Φ is the transition matrix associated with A. The matrix $W(t, \tau) = C(t)\Phi(t, \tau)B(\tau)$ will be called the *weighting pattern* of (2.1) [4], and a weighting pattern will be called *stationary* if $W(t, \tau) = W(t - \tau, 0)$. A weighting pattern W is called *realizable* if it can be realized by a finite-dimensional system (A, B, C), and the system (A, B, C) is called a *realization* of W. The system (2.1) is called *minimal* if there are no other realizations of W having a lower order.

DEFINITION 2.1. Two systems (A, B, C) and $(\hat{A}, \hat{B}, \hat{C})$ with corresponding state vectors x and $\hat{x}$ respectively are called *algebraically equivalent* whenever $\hat{x}(t) = T(t)x(t)$ for some absolutely continuous matrix function T possessing an absolutely continuous inverse. This equivalence will be denoted by $(A, B, C) \xrightarrow{T} (\hat{A}, \hat{B}, \hat{C})$.

If $(A, B, C) \xrightarrow{T} (\hat{A}, \hat{B}, \hat{C})$, then it is easily seen that $\hat{A}(t) = T(t)A(t)T^{-1}(t) + \dot{T}(t)T^{-1}(t)$, $\hat{B}(t) = T(t)B(t)$ and $\hat{C}(t) = C(t)T^{-1}(t)$. Also, if $\Phi(\cdot, \cdot)$ and $\hat{\Phi}(\cdot, \cdot)$ denote the transition matrices for A and $\hat{A}$ respectively, then $\hat{\Phi}(t, \tau) = T(t)\Phi(t, \tau)T^{-1}(\tau)$. Therefore, it is seen that two algebraically equivalent systems have the same weighting pattern.

For any given system (A, B, C), the operator δ is defined by (in any expression it will be tacitly assumed that A, B and C have the required number of derivatives)

$$(\delta C)(t) = \frac{d}{dt}C(t) + C(t)A(t) \tag{2.3}$$

and the operator Δ by

$$(\Delta B)(t) = -\frac{d}{dt}B(t) + A(t)B(t). \tag{2.4}$$

The powers δ^n and Δ^n are defined in the obvious way;

$$(\delta^n C)(t) = \frac{d}{dt}(\delta^{n-1}C)(t) + (\delta^{n-1}C)(t)A(t), \tag{2.5}$$

$$(\Delta^n B)(t) = -\frac{d}{dt}(\Delta^{n-1}B)(t) + A(t)(\Delta^{n-1}B)(t), \tag{2.6}$$

$$(\delta^0 C)(t) = C(t), \qquad (\Delta^0 B)(t) = B(t). \tag{2.7}$$

From these definitions, the "observability" and "controllability" matrices are defined respectively as

$$Q_n(t) = \begin{bmatrix} (\delta^0 C)(t) \\ (\delta^1 C)(t) \\ \vdots \\ (\delta^{n-1} C)(t) \end{bmatrix} \tag{2.8}$$

and

$$P_n(t) = [(\Delta^0 B)(t)|(\Delta^1 B)(t)| \cdots |(\Delta^{n-1} B)(t)]. \tag{2.9}$$

These matrices play a predominant role in specifying the conditions for controllability and observability of the system (A, B, C) (see [5], [6], [7]).

With regards to dynamic systems which realize a stationary weighting pattern, Silverman and Meadows have the following result.

THEOREM 2.1 (see [5]). *Suppose a system (A, B, C) is such that A and B are $n-1$ times continuously differentiable, and C is n times continuously differentiable. Then a necessary and sufficient condition for (A, B, C) to be a minimal realization of a stationary weighting pattern is that $\Gamma_{n+1,n}(t) = Q_{n+1}(t)P_n(t)$ be constant and have rank n.*

3. A condition for the realization of a stationary weighting pattern. The main result of this section is a slight improvement on the criterion given in Theorem 2.1. In particular, we have the following theorem.

THEOREM 3.1. *Suppose the system (A, B, C) is such that A is $2n-2$ times continuously differentiable, B and C are $2n-1$ times continuously differentiable. Then a necessary and sufficient condition for (A, B, C) to be a minimal realization of a stationary weighting pattern is that either $(\delta^k C)(t)B(t)$ or $C(t)(\Delta^k B)(t)$ be constant for $k = 0, 1, \cdots, 2n-1$, and the corresponding matrix*

$$\begin{bmatrix} (\delta^0 C)B & (\delta^1 C)B & \cdots & (\delta^{n-1} C)B \\ (\delta^1 C)B & (\delta^2 C)B & \cdots & (\delta^n C)B \\ \hline (\delta^{n-1} C)B & (\delta^n C)B & \cdots & (\delta^{2n-2} C)B \end{bmatrix}$$

or

$$\begin{bmatrix} C(\Delta^0 B) & C(\Delta^1 B) & \cdots & C(\Delta^{n-1} B) \\ C(\Delta^1 B) & C(\Delta^2 B) & \cdots & C(\Delta^n B) \\ \hline C(\Delta^{n-1} B) & C(\Delta^n B) & \cdots & C(\Delta^{2n-2} B) \end{bmatrix}$$

has rank n.

The improvement of this result over that of Theorem 2.1 is simply that only $2n$ matrix products must be checked for time invariance, whereas Theorem 2.1 requires $n(n-1)$ products to be checked.

The proof of Theorem 3.1 follows directly from Theorem 2.1 and the following lemma and its corollary.

LEMMA 3.1. *For a given system (A, B, C), if $(\delta^j C)(\Delta^k B)(t)$ is constant, then $(\delta^{j+1} C)(\Delta^k B) = (\delta^j C)(\Delta^{k+1} B)(t)$.*

Proof. Since $(\delta^i C)(\Delta^k B)$ is constant, its derivative is zero, which implies that

$$\frac{d}{dt}(\delta^j C)\,\Delta^k B = -\delta^j C \frac{d}{dt}(\Delta^k B). \tag{3.1}$$

But

$$(\delta^{j+1}C)(\Delta^k B) = \left[\frac{d}{dt}(\delta^j C) + (\delta^j C)A\right](\Delta^k B)$$
$$= \frac{d}{dt}(\delta^j C)\Delta^k B + (\delta^j C)A(\Delta^k B), \tag{3.2}$$

and using (3.1) in (3.2) gives

$$(\delta^{j+1}C)(\Delta^k B) = -(\delta^j C)\frac{d}{dt}(\Delta^k B) + (\delta^j C)A(\Delta^k B)$$
$$= (\delta^j C)(\Delta^{k+1}B). \tag{3.3}$$

COROLLARY 3.1. *Suppose* $(\delta^k C)(\Delta^0 B)$ *is constant for* $k = 0, 1, 2, \cdots, N$. *Then* $(\delta^i C)(\Delta^j B)$ *is constant for all* i *and* j *satisfying* $i + j \leqq N$. *Furthermore, if* $i_1 + j_1 = i_2 + j_2 \leqq N$, *then* $(\delta^{i_1}C)(\Delta^{j_1}B) = (\delta^{i_2}C)(\Delta^{j_2}B)$.

Proof. The proof proceeds by a repeated use of Lemma 3.2 and induction on $i + j$. First of all, by hypothesis, $(\delta^0 C)(\Delta^0 B)$ is constant, so the conclusion is true for $i + j = 0$. Suppose that $(\delta^i C)(\Delta^j B)$ is constant for all i and j satisfying $i + j = n < N$. Then, by Lemma 3.2,

$$(\delta^n C)(\Delta^0 B) \text{ const.} \Rightarrow (\delta^{n+1})(\Delta^0 B) = (\delta^n C)(\Delta^1 B),$$
$$(\delta^{n-1})(\Delta^0 B) \text{ const.} \Rightarrow (\delta^n C)(\Delta^1 B) = (\delta^{n-1})(\Delta^2 B),$$
$$\vdots$$
$$(\delta^1 C)(\Delta^{n-1} B) \text{ const.} \Rightarrow (\delta^2 C)(\Delta^{n-1} B) = (\delta^1 C)(\Delta^n B),$$
$$(\delta^0 C)(\Delta^n B) \text{ const.} \Rightarrow (\delta^1 C)(\Delta^{n+1} B) = (\delta^0 C)(\Delta^{n+1} B).$$

Hence, it is seen that

$$(\delta^{n+1}C)(\Delta^0 B) = (\delta^n C)(\Delta^1 B) = \cdots = (\delta^0 C)(\Delta^{n+1} B). \tag{3.4}$$

Since $n < N$, $n + 1 \leqq N$ and thus by hypothesis $(\delta^{n+1}C)(\Delta^0 B)$ is constant. Thus, from (3.4), it is seen that the conclusion is true for $i + j = n + 1$ if it is true for $i + j = n$.

By completely similar arguments it is easily seen that the following is also true.

COROLLARY 3.2. *Suppose* $(\delta^0 C)(\Delta^k B)$ *is constant for* $k = 0, 1, \cdots, N$. *Then the conclusion of Corollary* 3.1 *is true.*

Using Corollaries 3.1 and 3.2 in conjunction with Theorem 2.1 it is easily seen that Theorem 3.1 follows.

4. An application to feedback systems. Consider the feedback system shown in Fig. 1, in which L is a linear time-invariant system with representation

$$\dot{x}(t) = Ax(t) + Bu_1(t) + d\tilde{u}_2(t), \tag{4.1a}$$

$$y(t) = Cx(t), \tag{4.1b}$$

$$y_2(t) = fx(t). \tag{4.1c}$$

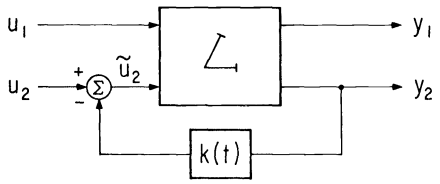


FIG 1. *Time-varying feedback system*

With the input $\tilde{u}_2(t) = u_2(t) - k(t)y_2(t)$, the dynamic equations for the feedback system are easily seen to be

$$\dot{x}(t) = (A - k(t)\,df)x(t) + Bu_1(t) + du_2(t), \tag{4.2a}$$

$$y_1(t) = Cx(t), \tag{4.2b}$$

$$y_2(t) = fx(t). \tag{4.2c}$$

The question is now raised as to whether the system of (4.2) can be a realization of a stationary weighting pattern if $dk/dt \not\equiv 0$. It will be shown that this can only happen in the trivial case when either d or f is the zero vector. This result is based on the following theorem.

THEOREM 4.1. *Consider the system* $(A + k(\cdot)df, B, C)$ *where* A *is a real constant* $n \times n$ *matrix,* d, f' *are real constant* n*-vectors,* B *and* C *are real constant* $n \times m$ *and* $p \times n$ *matrices, respectively, and* $k(\cdot)$ *is a* $2n - 2$ *times continuously differentiable real-valued function of* t *with* $dk/dt \not\equiv 0$. *Then if* $(A + k(\cdot)df, B, C)$ *realizes a stationary weighting pattern, the system* (A, B, C) *realizes the same weighting pattern. Furthermore, if* (A, B, C) *is minimal, then* $(A + k(\cdot)df, B, C)$ *realizes a stationary weighting pattern if and only if* $d = 0$ *or* $f = 0$.

Proof. The sufficiency of the last statement of the theorem is immediate since $d = 0$ or $f = 0$ results in the system (A, B, C).

The necessity will be proved by showing that the assumption that $(A + k(\cdot)df, B, C)$ is a realization of a stationary weighting pattern implies that one of the following two conditions must be satisfied:

(i) $CA^i d = 0$, $i = 0, 1, 2, \cdots, n - 1$;

(ii) $fA^i B = 0$, $i = 0, 1, 2, \cdots, n - 1$.

In order to prove (i) and (ii) the following lemma is needed.

LEMMA 4.1. *Consider the system* $(A + k(\cdot)df, b, c)$, *where* A, d, f *are as above and* b *and* c' *are real constant* n*-vectors. If* $cA^k d = 0$ *for* $k = 0, \cdots, N$, *then* $(\delta^k c)(t) = cA^k$ *for* $k = 0, 1, \cdots, N + 1$. *If* $fA^k b = 0$ *for* $k = 0, 1, \cdots, N$, *then* $(\Delta^k b)(t) = A^k b$ *for* $k = 0, 1, \cdots, N + 1$.

Proof. Suppose $cA^k d = 0$ for $k = 0, 1, \cdots, N$. The proof that $(\delta^j c)(t) = cA^j$ for $j = 0, 1, \cdots, N + 1$ will proceed by induction on j. First of all, $\delta^0 c = c$, so the result is true for $j = 0$. Suppose $(\delta^j c)(t) = cA^j$ for some $j \leqq N$. Then

$$(\delta^{j+1} c)(t) = \frac{d}{dt}(\delta^j c)(t) + (\delta^j c)(t)(A + k(t)df), \tag{4.3}$$

and since $(\delta^j c)(t) = cA^j$ and $cA^j d = 0$, (4.3) becomes

$$(\delta^{j+1} c)(t) = cA^{j+1} + k(t)cA^j df = cA^{j+1}, \tag{4.4}$$

which proves the result. A similar proof is used to show $\Delta^k b = A^k b$ for $k = 0, \cdots, N+1$ if $fA^k b = 0$ for $k = 0, 1, \cdots, N$.

Returning now to the proof of (i) and (ii), write B and C as follows:

$$B = (b_1, b_2, \cdots, b_m),$$

$$C' = (c_1', c_2', \cdots, c_p'),$$

where b_i and c_i' are constant n-vectors. Since $(A + k(\cdot)df, B, C)$ realizes a stationary weighting pattern, clearly every system $(A + k(\cdot)df, b_i, c_j)$ must also. Thus, from Theorem 3.1,[1] it is known that $(\delta^k c_j)b_i$ and $c_j(\Delta^k b_i)$ must be constant for $k = 0, 1, \cdots, 2n-1$. Since b_i and c_j are constant, these conditions are satisfied for $k = 0$. For $k = 1$ they give

$$[c_j A + k(t)c_j df]b_i = \text{const.} \tag{4.5}$$

and

$$c_j[Ab_i + k(t)df b_i] = \text{const.} \tag{4.6}$$

Since $dk/dt \not\equiv 0$, it is seen from (4.5) and (4.6) that $c_j df b_i = 0$, which leads to three possibilities:

(a) $c_j d = 0$ and $f b_i \neq 0$,
(b) $c_j d \neq 0$ and $f b_i = 0$,
(c) $c_j d = f b_i = 0$.

Suppose (a) holds and $c_j A^k d = 0$ for $k = 0, 1, \cdots, N-1$ $(N < 2n-2)$. Then, from Lemma 4.1, $(\delta^k c_j) = c_j A^k$ for $k = 0, 1, \cdots, N$. Hence

$$(\delta^{N+1} c_j)b_i = c_j A^{N+1} b_i + k(t)c_j A^N df b_i. \tag{4.7}$$

Using the fact that $(\delta^i c_j)b_i$ is constant for $i = 0, 1, \cdots, 2n-1$, it follows from (4.7) that $c_j A^N df b_i = 0$. But $f b_i \neq 0$, so $c_j A^N d = 0$. Consequently, by induction on k, it is seen that $c_j A^k d = 0$ for $k = 0, 1, \cdots, n-1$.

In a similar manner, (b) implies that $fA^k b_j = 0$ for $k = 0, 1, \cdots, n-1$. For case (c), let q be the first integer such that $c_j A^k d$ and $fA^k b_i$ do not both vanish. If there is no such integer $q \leqq n-1$, then both $c_j A^k d = 0$ and $fA^k b_i = 0$ for $k = 0, 1, \cdots, n-1$ and (i) and (ii) follow. Therefore, suppose $q < n-1$. From Corollary 3.1 of Lemma 3.1 and Theorem 3.1 it follows that $(\delta^m c_j)(\Delta^k b_i)$ is constant for $m + k \leqq 2n-1$. Hence, $(\delta^m c_j)(\Delta^q b_i)$ is constant for $m \leqq 2n-1-q$. Also, from Lemma 4.1 it is known that $c_j A^k d = fA^k b_i = 0$ for all $k < q$ implies that

$$\delta^k c_j = c_j A^k, \tag{4.8}$$

$$\Delta^k b_i = A^k b_i. \tag{4.9}$$

Hence,

$$(\delta^{q+1} c_j)(t)(\Delta^q b_i)(t) = [c_j A^{q+1} + k(t)c_j A^q df]A^q b_i, \tag{4.10}$$

and since $q + 1 < 2n - 1 - q$ (recall $q < n-1$), the left-hand side of (4.10) is constant. Thus there follows from (4.10),

$$(c_j A^q d)(fA^q b_i) = 0. \tag{4.11}$$

[1] If the minimality assumption is removed from Theorem 3.1, the necessity of $(\delta^k C)B = \text{const.}$ and $C(\Delta^k B) = \text{const.}$ remains while the remainder of the conclusions are no longer necessary (cf. [5]).

However, one of the terms in (4.11) is nonzero by the choice of q. Therefore, either

(a′) $c_jA^qd = 0$ and $fA^qb_i \neq 0$, or

(b′) $c_jA^qd \neq 0$ and $fA^qb_i = 0$.

It is easily seen that (a′) leads to $c_jA^kd = 0$ for $k = 0, 1, \cdots, n-1$, and (b′) leads to $fA^kb_i = 0$ for $k = 0, 1, \cdots, n-1$ in the same way as (a) and (b) did above. Thus, either $c_jA^kd = 0$ or $fA^kb_i = 0$ for $k = 0, 1, \cdots, n-1$. Suppose $c_jA^kd \neq 0$ for some j and $k \leqq n-1$, then $fA^kb_i = 0$ for all i and $k \leqq n-1$. Similarly, if $fA^kb_i \neq 0$ for some i and $k \leqq n-1$, then $c_jA^kd = 0$ for all j and $k \leqq n-1$. As a result, it is seen that either (i) or (ii) must hold.

If (A, B, C) is minimal, then the matrices $P = (B, AB, \cdots, A^{n-1}B)$ and $Q' = [C', A'C', \cdots, A'^{n-1}C']$ are of full rank. Thus, since (i) implies $Qd = 0$, it follows that $d = 0$ if (i) holds. On the other hand, if (ii) holds, then $fP = 0$ which implies that $f = 0$. Hence, if (A, B, C) is minimal and $(A + k(\cdot)df, B, C)$ realizes a stationary weighting pattern, then either $d = 0$ or $f = 0$.

Consider now the case when (A, B, C) is not minimal. It is easily seen that the zero state response of the system $(A + k(\cdot)df, B, C)$ is

$$y(t) = C\int_{t_0}^{t} e^{A(t-\tau)}[Bu(\tau) - k(\tau)dfx(\tau)]\,d\tau. \tag{4.12}$$

If $CA^kd = 0$ for $k = 0, 1, \cdots, n-1$, then it is easily seen that $Ce^{At}d \equiv 0$. Thus, if (i) holds, (4.12) becomes

$$y(t) = \int_{t_0}^{t} Ce^{A(t-\tau)}Bu(\tau)\,d\tau. \tag{4.13}$$

Now, suppose (ii) holds. With $x(t_0) = 0$ it is seen that

$$fx(t) = f\int_{t_0}^{t} e^{A(t-\tau)}[Bu(\tau) - k(\tau)dfx(\tau)]\,d\tau. \tag{4.14}$$

Since $fA^kB = 0$ for $k = 0, 1, \cdots, n-1$ implies that $fe^{At}B \equiv 0$, (4.14) becomes

$$fx(t) = -\int_{t_0}^{t} k(\tau)fdfx(\tau)\,d\tau. \tag{4.15}$$

Therefore, $fx(t)$ satisfies the differential equation

$$\frac{d}{dt}fx(t) = -k(t)fdfx(t). \tag{4.16}$$

and since $fx(t_0) = 0$ it follows that $fx(t) \equiv 0$. From (4.12) it is then seen that $y(t)$ is given by (4.13). Thus, also in this case it is seen that $(A + k(\cdot)df, B, C)$ has the same weighting pattern as the system (A, B, C). This completes the proof.

Returning now to the feedback system of Fig. 1, it is seen that the representation of (4.2) is in the form considered in Theorem 4.1. Since we are only concerned with input-output mappings, it is always possible to select a representation such that (4.2) satisfies the minimality assumption. Thus, it follows from Theorem 4.1 that the feedback system of Fig. 1 with $\dot{k} \not\equiv 0$ will realize a stationary weighting pattern only in the trivial cases of $d = 0$ or $f = 0$.

A less restrictive requirement on the system of Fig. 1 would be to set $u_2 \equiv 0$ and ask for a stationary weighting pattern for the input-output pair u_1 and y_1. For this case, the dynamic equations would be

$$\dot{x}(t) = (A - k(t)df)x(t) + Bu_1(t), \tag{4.17a}$$

$$y_1(t) = Cx(t). \tag{4.17b}$$

This system is also in the form considered in Theorem 4.1, although it need not be minimal. Nevertheless, Theorem 4.1 shows that if the weighting pattern of the system in (4.17) is stationary, then it is independent of the feedback $k(\cdot)$.

Acknowledgment. The author wishes to express his gratitude to Professor R. W. Brockett for many helpful discussions.

REFERENCES

[1] R. W. BROCKETT AND R. A. SKOOG, *Impedance synthesis using time varying resistors and capacitors*, Proc. 2nd Annual Princeton Conf. on Information Sciences and Systems, Princeton, N.J., 1968.

[2] R. A. SKOOG, *Time dependent realization theory with applications to electrical network synthesis*, Tech. Memorandum ESL-TM-393, M.I.T. Elect. Sys. Lab. 1969.

[3] R. W. BROCKETT AND R. A. SKOOG, *A new perturbation theory for nonlinear network synthesis*, Proc. of the 21st Annual Symposium on Applied Math., Durham, N.C., 1969.

[4] D. C. YOULA, *The synthesis of linear dynamical systems from prescribed weighting patterns*, SIAM J. Appl. Math., 14 (1966), pp.

[5] L. M. SILVERMAN AND H. E. MEADOWS, *Equivalent realizations of linear systems*, Ibid., 17 (1969), pp. 393–408.

[6] ———, *Controllability and observability in time-variable linear systems*, this Journal, 5 (1967), pp. 64–73.

[7] A. CHANG, *An algebraic characterization of controllability*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 112–113.

[8] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

ON A RECENT PROOF OF PONTRYAGIN'S NECESSARY CONDITIONS*

R. F. BAUM† AND L. CESARI‡

1. Introduction. In a recent paper [7] E. J. McShane has given an interesting proof of Pontryagin's necessary condition for generalized solutions. In his proof McShane assumes that the minimum is given by a generalized solution (defined by means of measures), and so corresponding solutions are also generalized ones.

We note in this paper that if McShane's proof is worded in terms of usual solutions only, with the assumption that the sets $Q(t, x) = f(t, x, U(t))$ are convex, then the proof becomes particularly straightforward, and a number of simplifications can be made. Also, by using this simplified proof, it is possible to prove Pontryagin's condition, not only in the usual (strong) form ((P_3), § 2), but also in a weakened form ((P_2), § 2) which we found of some relevance. Moreover, the same simplified proof for usual solutions holds as well for generalized solutions, when these are written as Gamkrelidze's chattering regimes. The latter are indeed still usual solutions with an augmented system of controls. Finally, the same proof with the simplifications mentioned above extends to the case where the points of the trajectory are elements of C (space of continuous functions) under a suitable hypothesis of convexity parallel to the one above. This situation occurs in a number of cases, as for instance in certain stochastic optimization problems discussed by R. F. Baum [3], [4], and in the optimization problems with delay recently discussed by T. S. Angell [1], precisely, problems monitored by functional equations (J. K. Hale [6]).

2. Description of the usual systems. We first consider ordinary control systems. Let A denote the constraint set, a closed subset of the tx-space $E_1 \times E_n$, with t in E_1, and $x = (x^1, \cdots, x^n)$, the space variable, in E_n. Let $U(t)$, the control set, be a subset of the u-space E_m, $u = (u^1, \cdots, u^m)$, the control variable. Let $M = \{(t, x, u) : (t, x) \in A, u \in U(t)\}$ be a closed subset of E_{1+n+m}, and let $f = (f_1, \cdots, f_n)$ be a continuous vector function from M into E_n. Let the boundary set B be a closed set of points (t_1, x_1, t_2, x_2) in E_{2n+2}, $x_1 = (x_1^1, \cdots, x_1^n)$, $x_2 = (x_2^1, \cdots, x_2^n)$. Let g be a continuous function from B into E_1.

We shall consider the class Ω of all pairs $x(t), u(t)$, $t_1 \leqq t \leqq t_2$, called admissible pairs, satisfying the following conditions:

(a) $x(t)$ is absolutely continuous in $[t_1, t_2]$;
(b) $u(t)$ is measurable in $[t_1, t_2]$;
(c) $(t, x(t)) \in A, t_1 \leqq t \leqq t_2$;
(d) $(t_1, x(t_1), t_2, x(t_2)) \in B$;
(e) $u(t) \in U(t)$ almost everywhere (a.e.) in $[t_1, t_2]$;
(f) the state equation $dx(t)/dt = f(t, x(t), u(t))$ is satisfied a.e. in $[t_1, t_2]$.

Let $\eta(x) = (t_1, x(t_1), t_2, x(t_2))$. The functional $I[x, u] = g(\eta(x)) = g(t_1, x(t_1), t_2, x(t_2))$ is called the cost functional.

* Received by the editors July 20, 1970, and in revised form August 11, 1970.
† Department of Industrial Engineering, University of Michigan, Ann Arbor, Michigan 48104.
‡ Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48104.

We seek the absolute minimum of $I[x,u]$ in the class Ω. If $(x_0,u_0)\in\Omega$ has the property that $I[x_0,u_0]\leqq I[x,u]$ for all $(x,u)\in\Omega$, then we say that x_0,u_0 is an optimal pair, and we may say that u_0 is an optimal control, and x_0 is an optimal trajectory. Though the optimal pair x_0,u_0 may not be unique in Ω, the value of the cost functional $I[x_0,u_0]$ is the same for all optimal pairs.

We now state necessary conditions for a pair $(x_0,u_0)\in\Omega$ to be an optimal pair.

THEOREM 1 (Pontryagin's necessary conditions). *Given a control system as described above, let us assume that $f(t,x,u)$ possesses continuous partial derivatives $f_t=(f_{it}=\partial f_i/\partial t,\ i=1,\cdots,n)$, $f_x=(f_{ix^j}=\partial f_i/\partial x^j, i,j=1,\cdots,n)$ in M, and that the set $Q(t,x)=f(t,x,U(t))=\{z\in E_n: z=f(t,x,u)$ for some u in $U(t)\}$ is convex in E_n for each (t,x) in A (see also Remark 2.1(a)).*

Let $x_0(t),u_0(t),t_1\leqq t\leqq t_2$, denote an optimal pair for which:

(i) *The graph of $x_0,[(t,x_0(t)),t_1\leqq t\leqq t_2]$, is interior to A.*

(ii) *$u_0(t)$ is bounded in $[t_1,t_2]$; that is, $|u_0(t)|\leqq d, t_1\leqq t\leqq t_2$, for some constant d (see Remark 2.1(b)).*

(iii) *The endpoint $\eta(x_0)=(t_1,x_0(t_1),t_2,x_0(t_2))$ of the optimal trajectory x_0 is a point of B, where B possesses a tangent hyperplane B' of some dimension k, $0\leqq k\leqq 2n+2$, whose vectors shall be denoted by $h=(\tau_1,\xi_1,\tau_2,\xi_2)$, with $\xi_1=(\xi_1^1,\cdots,\xi_1^n)$, $\xi_2=(\xi_2^1,\cdots,\xi_2^n)$, or in differential form, $h=(dt_1,dx_1,dt_2,dx_2)$, with $dx_1=(dx_1^1,\cdots,dx_1^n), dx_2=(dx_2^1,\cdots,dx_2^n)$.*

(iv) *g possesses a differential dg at $\eta(x_0)$, say*

$$dg=g_{t_1}\tau_1+\sum_{i=1}^{n}g_{x_1^i}\xi_1^i+g_{t_2}\tau_2+\sum_{i=1}^{n}g_{x_2^i}\xi_2^i$$

or

$$dg=g_{t_1}\,dt_1+\sum_{i=1}^{n}g_{x_1^i}\,dx_1^i+g_{t_2}\,dt_2+\sum_{i=1}^{n}g_{x_2^i}\,dx_2^i,$$

where $g_{t_1},\cdots,g_{x_2^n}$ denote partial derivatives of g with respect to $t_1,\cdots,x_2^n$, all computed at $\eta(x_0)$.

Let the Hamiltonian H be defined by:

$$H(t,x,u,\lambda)=\lambda f(t,x,u)=\lambda_1 f_1+\cdots+\lambda_n f_n.$$

Then there exists a family of vector functions

$$\lambda(t)=(\lambda_1(t),\cdots,\lambda_n(t)),\qquad t_1\leqq t\leqq t_2,$$

which we shall call multipliers, such that:

(P_1) *$\lambda(t)$ is absolutely continuous in $[t_1,t_2]$, and satisfies*

$$d\lambda_i(t)/dt=-H_{x^i}(t,x_0(t),u_0(t),\lambda(t)),$$

$i=1,2,\cdots,n$, for almost all (a.a.) t in $[t_1,t_2]$. If dg is not identically zero at $\eta(x_0)$, then $\lambda(t)$ is never zero in $[t_1,t_2]$.

(P_2) (Weak minimum principle). *Given any bounded, measurable function $u(t), u(t)\in U(t)$ a.e. in $[t_1,t_2]$, then for a.a. t in $[t_1,t_2]$, $H(t,x_0(t),u_0(t),\lambda(t))\leqq H(t,x_0(t),u(t),\lambda(t))$.*

(P_3) (Usual minimum principle). *Let $U(t)=U, t_1\leqq t\leqq t_2$, be a fixed closed subset of E_m (see also Remark 5.1). Then for a.a. fixed t in $[t_1,t_2]$, $M(t)=M(t,x_0(t),\lambda(t))=H(t,x_0(t),u_0(t),\lambda(t))$ for a.a. t in $[t_1,t_2]$, where $M(t,x,\lambda)$ is defined by*

$$M(t,x,\lambda)=\inf_{u\in U(t)}H(t,x,u,\lambda),\qquad (t,x,\lambda)\in A\times E_n.$$

(P_4) *The function* $M(t) = M(t, x_0(t), \lambda(t))$ *is absolutely continuous in* $[t_1, t_2]$, *and*

$$\frac{dM(t)}{dt} = \frac{d}{dt}M(t, x_0(t), \lambda(t)) = -H_t(t, x_0(t), u_0(t), \lambda(t))$$

for a.a. t *in* $[t_1, t_2]$.

(P_5) (Transversality relation). *There is a constant* $\lambda_0 \geqq 0$ *such that*

$$(\lambda_0 g_{t_1} - M(t_1))\,dt_1 + \sum_{i=1}^{n} (\lambda_0 g_{x_1^i} + \lambda_i(t_1))\,dx_1^i$$

$$+ (\lambda_0 g_{t_2} + M(t_2))\,dt_2 + \sum_{i=1}^{n} (\lambda_0 g_{x_2^i} - \lambda_i(t_2))\,dx_2^i = 0$$

for every vector $H = (dt_1, dx_1, dt_2, dx_2)$ *in* B'.

We shall prove these necessary conditions under the simplifying assumption that t_1, t_2 are fixed. This proof is based on that of McShane [7]. For the extension of this proof with this assumption removed, see [5], [7] or [10].

Remark 2.1. (a) As usual, we may remove the assumption that $Q(t, x)$ is convex if we require $U(t) = U$, a fixed subset of E_m, and then use the usual "needle-like" variations in the proof of the necessary conditions (see [10]).

(b) We may weaken assumption (ii) together with the assumption that $U(t)$ must be closed. See [5] and [7] for details.

3. Variations. As in most proofs of Pontryagin's necessary conditions, we introduce a class of variations such that the corresponding trajectories satisfy exactly the differential system, the constraints, and the initial conditions, but not necessarily the terminal boundary conditions.

By a *variation* we shall denote a triple $v = (c, u, h)$ made up of a nonnegative number c, a (bounded) measurable m-vector function $u(t) = (u^1, \cdots, u^m)$, $t_1 \leqq t \leqq t_2$, with $u(t) \in U(t)$ for almost all $t \in [t_1, t_2]$, and a $(2n + 2)$-vector $h = (\tau_1, \xi_1, \tau_2, \xi_2) \in B'$, that is, a tangent vector to B at the point $\eta(x_0)$. Here $\eta(x_0) = (t_1, x_0(t_1), t_2, x_0(t_2))$, where t_1, t_2 are fixed; hence $\tau_1 = \tau_2 = 0$. We can think of $h = (0, \xi_1, 0, \xi_2)$ as the vector tangent to a curve C of class C^1 lying in B and issued from $\eta(x_0)$, say in this case C: $(t_1, x_1(z), t_2, x_2(z))$, $0 \leqq z \leqq 1$, and hence

$$(t_1, x_1(z), t_2, x_2(z)) \in B, \qquad 0 \leqq z \leqq 1,$$

$$x_0(t_1) = X_1(0), \quad x_0(t_2) = X_2(0), \quad \xi_1 = X_1'(0), \quad \xi_2 = X_2'(0),$$

where $X_1(z), X_2(z)$ are continuously differentiable in $[0, 1]$. We extend $X_1(z)$, $X_2(z)$ in the whole interval $[-1, 1]$ so that $X_1(z)$, $X_2(z)$ are still continuously differentiable in the whole interval $[-1, 1]$ (without requesting that the added arc of the curve C lie in B, though this may well be the case if say B is a smooth manifold in a neighborhood of $\eta(x_0)$). With this extension we can say that ξ_1 is the tangent to the curve C': $x = X_1(z)$ at $z = 0$, and ξ_2 is the tangent to the curve C'': $x = X_2(z)$ at $z = 0$.

Let us consider now an arbitrary system of s variations as before, say $v_\sigma = (c_\sigma, u_\sigma, h_\sigma)$, $\sigma = 1, \cdots, s$. Then $h_\sigma = (0, \xi_{1\sigma}, 0, \xi_{2\sigma}) \in B'$, $\sigma = 1, \cdots, s$, and each h_σ is a tangent vector to B at $\eta(x_0)$. As before, we can think of each h_σ as the

tangent vector to a curve C_σ of class C^1 lying in B and issued from $\eta(x_0), \sigma = 1, \cdots, s$. Nevertheless, it is convenient to think of those s curves C_σ as belonging to a suitable manifold of dimension s lying in B. In other words, we introduce the vector variable $z = (z_1, \cdots, z_s)$ varying in the interval $I = [z : 0 \leqq z_\sigma \leqq 1, \sigma = 1, \cdots, s]$, and vector functions $X_1(z), X_2(z)$ defined in I such that

$$(3.1) \qquad \begin{gathered} (t_1, X_1(z), t_2, X_2(z)) \in B \quad \text{for } z \in I, \\ x_0(t_1) = X_1(0), \quad x_0(t_2) = X_2(0), \quad \xi_{1\sigma} = X_{1z_\sigma}(0), \quad \xi_{2\sigma} = X_{2z_\sigma}(0), \end{gathered}$$

where $X_1(z), X_2(z)$ are continuously differentiable in I, and $X_{1z_\sigma}(0), X_{2z_\sigma}(0)$ denote the partial derivatives of X_1, X_2 with respect to z_σ at $0 = (0, \cdots, 0)$. Note that X_1, X_2 represent the curve C_σ above when z describes the interval $0 \leqq z_\sigma \leqq 1$ of the z_σ-axis, $\sigma = 1, \cdots, s$. As before, we extend the function $X_1(z), X_2(z)$ in the whole interval $V = [z : -1 \leqq z_\sigma \leqq 1, \sigma = 1, \cdots, s]$ so that $X_1(z), X_2(z)$ are still continuously differentiable in the whole V (without requesting that the new parts of the manifold so added lie in B). For every $z \in V$ we consider now the differential system with initial conditions

$$(3.2) \qquad dx/dt = q_z(t, x), \quad t_1 \leqq t \leqq t_2, \quad x(t_1, z) = X_1(z),$$

where $q_z(t, x)$ is a vector function of t and x, depending on the parameter $z \in V$, defined by

$$(3.3) \qquad q_z(t, x) = (1 - \sum_\sigma c_\sigma z_\sigma) f(t, x, u_0(t)) + \sum_\sigma c_\sigma z_\sigma f(t, x, u_\sigma(t)),$$

$t_1 \leqq t \leqq t_2$. Note that for $z = 0 = (0, \cdots, 0)$ we have $q_0(t, x) = f(t, x, u_0(t))$, $X_1(0) = x(t_1, 0) = x_0(t_1)$. Thus for $z = 0$, $x(t, 0)$ satisfies the same differential system and initial conditions as $x_0(t)$, and hence, by the usual uniqueness theorems (see [8]), $x(t, 0) = x_0(t)$ for all $t_1 \leqq t \leqq t_2$.

The graph $[(t, x_0(t)), t_1 \leqq t \leqq t_2]$ of x_0 has been assumed to lie in the interior of A. Thus, there is some $\delta > 0$ such that (t, x) is certainly in A if $t_1 \leqq t \leqq t_2$, and $|x - x_0(t)| \leqq \delta$. We have assumed $u_0(t), u_1(t), \cdots, u_s(t)$ all bounded, say in absolute value $\leqq M'$, and f is continuous. Thus for $t_1 \leqq t \leqq t_2$, $|x - x_0(t)| \leqq \delta, |u| \leqq M'$, $u \in U(t), f(t, x, u)$ is bounded, say $|f(t, x, u)| \leqq M''$, and hence $q_z(t, x)$ is also bounded, say $|q_z(t, x)| \leqq M'''$ for the same t, x and $z \in V$. Finally, $X_1(z)$ is a continuous function of z. We can hence conclude (see [8]), that there is some number $\gamma, 0 < \gamma \leqq 1$, such that for $|z_\sigma| \leqq \gamma$ the solution $x(t, z)$ of the differential system and initial condition (3.2) exists in the whole interval $[t_1, t_2]$ and $|x(t, z) - x(t, 0)| = |x(t, z) - x_0(t)| \leqq \delta$, $t_1 \leqq t \leqq t_2$. Thus, if we denote by V_γ the interval $V_\gamma = [-\gamma \leqq \xi_\sigma \leqq \gamma, \sigma = 1, \cdots, s]$, we conclude that for every $z \in V_\gamma$ certainly $x(t, z)$ exists in the whole interval $[t_1, t_2]$ and its graph lies in A. For $z \in V_\gamma \cap I$ and again $\gamma > 0$ sufficiently small, we have $c_\sigma z_\sigma \geqq 0$, $\sigma = 1, \cdots, s$, $1 - \sum_\sigma c_\sigma z_\sigma \geqq 0$; and hence $q_z(t, z)$ is a convex combination of the $s + 1$ points $f(t, x, u_0(t)), f(t, x, u_\sigma(t)), \sigma = 1, \cdots, s$, all in the convex subset $Q(t, x)$ of E_n. Hence $q_z(t, z) \in Q(t, x)$ or $dx(t, z)/dt \in q_z(t, x(t, z))$ for almost all $t \in [t_1, t_2]$ and $z \in V_\gamma \cap I$. Thus by force of the implicit function theorem for orientor fields (see [9]) and $z \in V_\gamma \cap I$, there is a measurable vector function $u(t, z), t_1 \leqq t \leqq t_2$, such that $u(t, z) \in U(t)$ and $dx(t, z)/dt = f(t, x(t, z)),$

$u(t, z)$) for almost all $t \in [t_1, t_2]$, that is,

$$dx(t, z)/dt = f(t, x(t, z), u(t, z)) \tag{3.4}$$

a.e. in $[t_1, t_2]$. From (3.2) we can also conclude (see [8]) that for $z \in V_\gamma$, the solution $x(t, z)$ possesses partial derivatives $\partial x(t, z)/\partial z_\sigma$, $\sigma = 1, \cdots, s$, with respect to each z_σ, and that these derivatives satisfy

$$\begin{aligned}(d/dt)(\partial x(t, z)/\partial z_\sigma) = &-c_\sigma f(t, z, u_0(t)) + c_\sigma f(t, x(t, z), u_\sigma(t)) \\ &+ (1 - cz) f_x(t, x(t, z), u_0(t))(\partial x(t, z)/\partial z_\sigma) \\ &+ (cz) f_x(t, x(t, z), u_\sigma(t))(\partial x(t, z)/\partial z_\sigma), \\ (\partial x(t, z)/\partial z_\sigma)_{(t=t_1)} = &\ X_{1z_\sigma}(z), \qquad \sigma = 1, \cdots, s,\end{aligned} \tag{3.5}$$

where f_x denotes the $n \times n$ matrix $(f_{ix^j}, i, j = 1, \cdots, n)$. We are interested in these partial derivatives at $z = 0$, that is, we are interested in the functions of t defined by

$$y_\sigma(t) = [\partial x(t, z)/\partial z_\sigma]_{z=0}, \qquad t_1 \leqq t \leqq t_2, \qquad \sigma = 1, \cdots, s,$$

when $y(t) = (y^1, \cdots, y^n)$. Relations (3.5) for $z = 0$ yield immediately

$$\begin{aligned}dy_\sigma/dt = &-c_\sigma f(t, x_0(t), u_0(t)) + c_\sigma f(t, x_0(t), u_\sigma(t)) \\ &+ f_x(t, x_0(t), u_0(t)) y_\sigma(t), \qquad t_1 \leqq t \leqq t_2, \\ y_\sigma(t_1) = &\ X_{1z_\sigma}(0) = \xi_{1\sigma}, \qquad \sigma = 1, \cdots, s.\end{aligned} \tag{3.6}$$

Note that $y_\sigma(t)$ depends only on the variation $v_\sigma = (c_\sigma, u_\sigma, h_\sigma)$, and we shall often denote it by $y(t; v_\sigma)$. Equation (3.6) is the usual *variational equation* (with its initial data) of $x_0(t)$ with respect to the variation $v_\sigma = (c_\sigma, u_\sigma, h_\sigma)$, $h_\sigma = (0, \xi_{1\sigma}, 0, \xi_{2\sigma})$. Note that the equation and initial data (3.6) determine $y_\sigma(t)$, that is, $y(t; v_\sigma)$, uniquely in the whole interval $[t_1, t_2]$.

4. The cone K. We shall now consider the cone K in E_{n+1} made up of the terminal points of the linearized trajectories in E_n corresponding to all possible variations defined above, and associated values of the linearized cost functional. We shall prove that the point $(-1, 0, \cdots, 0)$ in E_{n+1} is not an interior point of K. The argument is by contradiction, showing that in the opposite case, there would exist an admissible trajectory giving a lower cost than the optimal cost. The proof of the Pontryagin necessary conditions then follows by taking a supporting hyperplane to K.

For every variation $v = (c, u, h)$ with $h = (0, \xi_1, 0, \xi_2)$ let us consider the $(n + 1)$-vector $\tilde{Y}(v) = (Y^0, Y) = (Y^0, Y^1, \cdots, Y^n)$ defined by $Y^0(v) = g_{x_1}\xi_1 + g_{x_2}\xi_2$, $Y(v) = y(t_2; v) - \xi_2$, where g_{x_1}, g_{x_2} are the $1 \times n$ matrices of the partial derivatives of $g(t_1, x_1, t_2, x_2)$ with respect to the arguments $x_1^1, \cdots, x_1^n$, or $x_2^1, \cdots, x_2^n$ respectively, these partial derivatives being evaluated at the point $\eta(x_0) = (t_1, x_0(t_1), t_2, x_0(t_2))$, and where $y(t; v)$ simply denotes the solution of the differential system and initial data (3.6) with $v_\sigma = v = (c, u, h)$, $h = (0, \xi_1, 0, \xi_2)$. We shall denote by $K \subset E_{n+1}$ the set of all such vectors $\tilde{Y}(v)$ in $z^0 z^1 \cdots z^n$-space $E_n + 1$.

LEMMA 4.1. *K is a convex cone with vertex at $(0, \cdots, 0)$; that is, if $\tilde{Y}(V_1)$, $\tilde{Y}(V_2) \in K$, and $a_1, a_2 \geqq 0$, then there is a variation $v = (c, y, h)$, $h = (0, \xi_1, 0, \xi_2)$ such that $\tilde{Y}(v) = a_1 Y(v_1) + a_2 Y(v_2)$.*

Proof. Let $v_\sigma = (c_\sigma, u_\sigma, h_\sigma)$, $h_\sigma = (0, \xi_{1\sigma}, 0, \xi_{2\sigma})$, $\sigma = 1, 2$, be two given variations and $\tilde{Y}(v_1)$, $\tilde{Y}(v_2)$ the corresponding vectors in E_{n+1}. Assume first $a_1c_1 + a_2c_2 \neq 0$, and hence, since $a_1, a_2, c_1, c_2 \geqq 0$, $a_1c_1 + a_2c_2 > 0$. Take $h = a_1h_1 + a_2h_2$, $c = a_1c_1 + a_2c_2$; hence, if $h = (0, \xi_1, 0, \xi_2)$, then $\xi_\sigma = a_1\xi_{1\sigma} + a_2\xi_{2\sigma}$, $\sigma = 1, 2$. Let us consider the convex combination of $f(t, x_0(t), u_\sigma(t))$, $\sigma = 1, 2$,

$$q(t) = (a_1c_1 + a_2c_2)^{-1}[a_1c_1 f(t, x_0(t), u_1(t)) + a_2c_2 f(t, x_0(t), u_2(t))]. \tag{4.1}$$

Since $f(t, x_0(t), u_\sigma(t)) \in Q(t, x_0(t))$, $\sigma = 1, 2$, and $Q(t, x_0(t))$ is convex, we see that $q(t) \in Q(t, x_0(t))$ for almost all $t \in [t_1, t_2]$. By the implicit function theorem for orientor fields there is a measurable control function $u(t)$, $t_1 \leqq t \leqq t_2$, $u(t) \in U(t)$, such that

$$q(t) = f(t, x_0(t), u(t)) \quad \text{a.e. in } [t_1, t_2]. \tag{4.2}$$

We denote by v the new variation $v = (c, u, h)$ with $h = (0, \xi_1, 0, \xi_2)$. If $y(t; v_\sigma)$, $\sigma = 1, 2$, denote the solutions of the variational equation and initial data (3.6) relative to the variation v_σ, and y denotes the linear combination $y(t) = a_1 y(t; v_1) + a_2 y(t; v_2)$, then by linear combination of the relevant equations (4.2) with coefficients a_1, a_2 and the use of (4.1), (3.6) and of the definitions of c, h, we obtain for y the equation and initial data

$$\begin{aligned} dy/dt &= -cf(t, x_0(t), u_0(t)) + cf(t, x_0(t), u(t)) \\ &\quad + f_x(t, x_0(t), u_0(t))y(t), \qquad t_1 \leqq t \leqq t_2, \\ y(t_1) &= \xi_1 ; \end{aligned} \tag{4.3}$$

that is, $y(t) = y(t; v)$ is the unique solution of (3.6) relative to the new variation v.

From (4.3) we obtain:

$$\begin{aligned} a_1 Y^0(v_1) + a_2 Y^0(v_2) &= a_1(g_{x_1}\xi_{11} + g_{x_2}\xi_{21}) + a_1(g_{x_1}\xi_{12} + g_{x_2}\xi_{22}) \\ &= g_{x_1}\xi_1 + g_{x_2}\xi_2 = Y^0(v), \\ a_1 Y(v_1) + a_2 Y(v_2) &= a_1(y(t_2; v_1) - \xi_{21}) + a_2(y(t_2; v_2) - \xi_{22}) \\ &= y(t_2; v) - \xi_2 = Y(v), \end{aligned}$$

or $a_1\tilde{Y}(v_1) + a_2\tilde{Y}(v_2) = \tilde{Y}(v)$. If $a_1c_1 + a_2c_2 = 0$, then $a_1c_1 = a_2c_2 = 0$, and (4.3) become $dy/dt = f_x(t, x_0(t), u_0(t))$, $y(t_1) = \xi_1$, $t_1 \leqq t \leqq t_2$, and hence, the above argument holds for the variation $v = (0, u_0, h)$. We have thus proved that K is a convex cone.

LEMMA 4.2. *The point $(-1, 0, \cdots, 0)$ is not interior to K.*

Proof. Assume, if possible, that $(-1, 0, \cdots, 0)$ is interior to K. Then for some $\delta > 0$ sufficiently small the $n + 1$ points in E_{n+1}

$$\begin{gathered} (-1, -\delta, 0, \cdots, 0), (-1, 0, -\delta, 0, \cdots, 0), \cdots, (-1, 0, \cdots, 0, -\delta), \\ (-1, \delta, \delta, \cdots, \delta) \end{gathered} \tag{4.4}$$

certainly belong to K and hence there are variations $v_1, v_2, \cdots, v_{n+1}$ such that the

corresponding vectors $\tilde{Y}(v_1), \cdots, \tilde{Y}(v_{n+1})$ are exactly the corresponding vectors (4.4) with $v_\sigma = (c_\sigma, u_\sigma, h_\sigma), h_\sigma = (0, \xi_{1\sigma}, 0, \xi_{2\sigma}), \sigma = 1, \cdots, n+1$. We shall now take $s = n+1$ at the beginning of this proof and denote by $X_1(z)$, $X_2(z)$, $z = (z_1, \cdots, z_{n+1}) \in V$, the corresponding functions, and by $x(t, z) = (x^1, \cdots, x^n)$ the corresponding solution of (3.3) with initial values $x(t_1, z) = X_1(z)$. We have now to compare the end values of $x(t, z)$ or $x(t_2, z)$ with $X_2(z)$, and the value of the functional $g(\eta(x(t, z))$ with $g(\eta(x_0)) = g_{\min}$. Namely we shall consider the equations

$$\begin{aligned} g(t_1, X_1(z), t_2, X_2(z)) + z_0 - g_{\min} &= 0, \\ x^i(t_2, z) - X_2^i(z) &= 0, \qquad i = 1, \cdots, n. \end{aligned} \tag{4.5}$$

These $n+1$ equations (in the $n+2$ unknowns $z_1, \cdots, z_{n+1}, z_0$) are obviously satisfied by $z_1 = \cdots = z_{n+1} = z_0 = 0$ since then $x(t_1, 0) = x_0(t_1) = X_1(0)$, $x(t_2, 0) = x_0(t_2) = X_2(0)$, and $g(\eta(x_0)) = g_{\min}$. At the point $(0, \cdots, 0, 0)$ the partial derivatives of the first members with respect to, say, z_σ are respectively

$$\sum_{j=1}^{n} (g_{x_1^j} X_{1z_\sigma}^j + g_{x_2^j} X_{2z_\sigma}^j)_{z=0} = \sum_{j=1}^{n} (g_{x_1^j} \xi_{1\sigma}^j + g_{x_2^j} \xi_{2\sigma}^j) = Y^0(v_\sigma),$$

$$(\partial x^i(t_2, z)/\partial z_\sigma)_{z=0} - (X_{2z_\sigma}^i(z))_{z=0} = y^i(t_2; v_\sigma) - \xi_2^i = Y^i(v_\sigma),$$

$$i = 1, \cdots, n, \quad \sigma = 1, \cdots, n+1.$$

In other words, the $(n+1) \times (n+1)$ functional determinant of the $n+1$ equations (4.5) with respect to the $n+1$ variables $z_1, z_2, \cdots, z_{n+1}$ is the determinant of the $n+1$ vectors (4.4), and this determinant is $(-1)^{n+1}(n+1)\delta^n \neq 0$. By the implicit function theorem of calculus we conclude that for every $z_0 \neq 0$ and sufficiently small, the $n+1$ equations (4.5) can be solved with respect to $z_1, \cdots, z_{n+1}$, and that again for z_0 sufficiently small, the solutions

$$z_\sigma = Z_\sigma(z_0), \quad \sigma = 1, 2, \cdots, n+1, \quad \text{or} \quad Z(z_0) = (Z_1, \cdots, Z_{n+1})$$

are continuously differentiable functions of z_0. In other words, there is a neighborhood $(-\lambda, \lambda)$ of $z_0 = 0$ such that for $-\lambda \leqq z_0 \leqq \lambda$, the functions $Z_\sigma(z_0)$ satisfy the equations

$$\begin{aligned} g(t_1, X_1(Z(z_0)), t_2, X_2(Z(z_0))) + z_0 - g_{\min} &= 0, \\ x^i(t_2, Z(z_0)) - X_2^i(Z(z_0)) &= 0, \qquad i = 1, \cdots, n. \end{aligned} \tag{4.6}$$

Note that we have also

$$x^i(t_1, Z(z_0)) - X_1^i(Z(z_0)) = 0, \qquad i = 1, \cdots, n,$$

since this relation holds for all z as stated in (3.2). Thus, for every $z = Z(z_0)$, $-\lambda_0 \leqq z_0 \leqq \lambda_0$, we have

$$(t_1, x(t_1, z), t_2, x(t_2, z)) = (t_1, X_1(z), t_2, X_2(z)) \in B. \tag{4.7}$$

Again, by the implicit function theorem, the derivatives of the functions $Z_\sigma(z_0)$, $\sigma = 1, \cdots, n+1$, at $z_0 = 0$ can be obtained by differentiating relations (4.6) with

respect to z_0 and taking $z_0 = 0$. We obtain

$$\sum_{\sigma=1}^{n+1} Y_\sigma^0(v_\sigma)Z'_\sigma(0) + 1 = 0, \qquad \sum_{\sigma=1}^{n+1} Y^i(v_\sigma)Z'_\sigma(0) = 0, \qquad i = 1, \cdots, n, \tag{4.8}$$

where the coefficients of this system are given by relations (4.4), that is,

$$-Z'_1(0) - Z'_2(0) - \cdots - Z'_{n+1}(0) + 1 = 0,$$

$$-\delta Z'_1(0) + \delta Z'_{n+1}(0) = 0, \cdots, -\delta Z'_n(0) + \delta Z'_{n+1}(0) = 0;$$

and hence,

$$Z'_1(0) = Z'_2(0) = \cdots = Z'_{n+1}(0) = (n+1)^{-1} > 0. \tag{4.9}$$

Since the matrix of the coefficients of (4.8) is nonsingular, this is the only solution. We conclude that for z_0 positive and sufficiently small, say again $0 < z_0 \leqq \lambda$, the numbers $z_\sigma = Z_\sigma(z_0)$, $\sigma = 1, \cdots, n+1$, are all positive and as close to zero as we want since $Z_\sigma(0) = 0$. Thus $z = Z(z_0) = (Z_1, \cdots, Z_{n+1}) \in V_\sigma \cap I$ for $0 < z_0 \leqq \lambda$. From (3.4) and (4.6) we conclude that, for $0 < z \leqq \lambda$ and $z = Z(z_0)$, the pair $x(t, z), u(t, z), t_1 \leqq t \leqq t_2$, is admissible. From (4.9) we now have

$$I[x(t, z), u(t, z)] = \eta(x(t, z)) = g_{\min} - z_0 < g_{\min} \quad \text{for } 0 < z_0 \leqq \lambda,$$

and this contradicts the definition of $g_{\min}$. We have proved that $(-1, 0, \cdots, 0)$ is not an interior point of K. This completes the proof of Lemma 4.2.

LEMMA 4.3. *There are numbers* $\chi_0, \chi_1, \cdots, \chi_n$ *not all zero,* $\chi_0 \geqq 0$, *such that* $\sum_{i=0}^n \chi_i Y^i(v) \geqq 0$ *for all variations* v.

Proof. If K has no interior points, then Lemma 4.3 follows immediately. If K has interior points, then by Lemma 4.1 and Lemma 4.2, K possesses a supporting hyperplane through $(0, \cdots, 0)$, say $\sum_{i=0}^n \chi_i z^i = 0$, with K contained in $\sum_{i=0}^n \chi_i z^i \geqq 0$, and $(-1, 0, \cdots, 0)$ contained in $\sum_{i=0}^n x_i z^i \leqq 0$; in particular $\chi_0 \geqq 0$. Lemma 4.3 is thereby proved.

Remark 4.1. Note that, whenever $\chi_1 = \cdots = \chi_n = 0$, then $\chi_0 > 0$, the hyperplane becomes $z^0 = 0$, and K is contained in the half-space $z^0 \geqq 0$, or $Y^0(v) \geqq 0$ for all variations v. If we assume that g possesses a differential dg at $\eta(x_0)$, then Y^0 is this differential dg, a linear function, and $dg \geqq 0$ then implies $dg = 0$ (identically). Thus for dg not identically zero at $\eta(x_0)$, the n-vector $(\chi_1, \cdots, \chi_n)$ must be nonzero.

5. Completion of proof.

Proof of (P_1). Given any variation $v = (c, u, h)$ with $h = (0, \xi_1, 0, \xi_2)$, the corresponding variational equation and initial data (3.6) can be written in the form

$$dy/dt = A(t)y + b(t), \qquad y(t_1) = \xi_1, \tag{5.1}$$

where A is the $n \times n$ matrix $A(t) = f_x(t, x_0(t), u_0(t))$, and b is the n-vector (or $n \times 1$) matrix

$$b(t) = -cf(t, x_0(t), u_0(t)) + cf(t, x_0(t), u(t)).$$

We shall denote by χ the constant n-vector $\chi = (\chi_1, \cdots, \chi_n)$. We shall denote by $\Phi(t) = (\phi_{ij}, i, j = 1, \cdots, n)$, $t_1 \leqq t \leqq t_2$, any fundamental system of solutions of the homogeneous linear system $dy/dt = A(t)y$, $t_1 \leqq t \leqq t_2$, and by $\Phi^{-1}(t)$

$= (\psi_{ij}, i,j = 1, \cdots, n)$ the inverse matrix of Φ. We shall also denote by A_{-1} the transpose of any matrix A. With these notations the solution $y(t;v)$, $t_1 \leqq t \leqq t_2$, of the variational equation and initial data (5.1) can be written in the explicit form

$$y(t;v) = \Phi(t)[\Phi^{-1}(t_1)\xi_1 + \int_{t_1}^{t} \Phi^{-1}(\tau)[-cf(\tau, x_0(\tau), u_0(\tau)) + cf(\tau, x_0(\tau), u(\tau))]\, d\tau. \tag{5.2}$$

We shall define λ_0 and the n-vector $\lambda(t) = (\lambda_1, \cdots, \lambda_n)$ of the multipliers by taking

$$\lambda_0(t) = \lambda_0 = \chi_0, \qquad \lambda(t) = (\chi_{-1}\Phi(t_2)\Phi^{-1}(t))_{-1}. \tag{5.3}$$

Thus, we have immediately $\lambda(t_2) = \chi$, and $\lambda(t)$ is absolutely continuous in $[t_1, t_2]$. By direct differentiation, it follows that $d\lambda/dt = -(A(t))_{-1}\lambda(t)$; from this and Remark 4.1, (P_1) follows.

Proof of (P_5). For $a, b \in E_n$, let $a \cdot b = \sum_{i=1}^{n} a_i b_i$ be the usual inner product. Let v denote any variation $v = (c, u, h)$, $h = (0, \xi_1, 0, \xi_2)$. Let us replace $y(t_2;v)$ by its expression (5.2) in Lemma 4.3. We obtain

$$\lambda_0(g_{x_1}\xi_1 + g_{x_2}\xi_2) + \chi \cdot \Phi(t_2)\Big\{\Phi^{-1}(t_1)\xi_1 + \int_{t_1}^{t_2} \Phi^{-1}(\tau)[-cf(\tau, x_0(\tau), u_0(\tau)) + cf(\tau, x_0(\tau), u(\tau))]\, d\tau - \xi_2\Big\} \geqq 0,$$

where

$$\chi \cdot \Phi(t_2)\Phi^{-1}(t_1)\xi_1 = \chi_{-1}(\Phi(t_2)\Phi^{-1}(t_1)\xi_1) = (\chi_{-1}\Phi(t_2)\Phi^{-1}(t_1))\xi_1 = (\lambda(t_2))_{-1}\xi_1 = \lambda(t_2) \cdot \xi_1,$$

and analogously,

$$\chi \cdot \Phi(t_2)\Phi^{-1}(\tau)f = \lambda(\tau) \cdot f.$$

Hence,

$$\lambda_0(g_{x_1}\xi_1 + g_{x_2}\xi_2) - \lambda(t_2) \cdot \xi_2 + \lambda(t_1) \cdot \xi_1 + \int_{t_1}^{t_2} [-c\lambda(\tau) \cdot f(\tau, x_0(\tau), u_0(\tau)) + c\lambda(\tau) \cdot f(\tau, x_0(\tau), u(\tau))]\, d\tau \geqq 0,$$

and using the definition of the Hamiltonian we obtain

$$\sum_{i=1}^{n} [\lambda_0 g_{x_1^i} + \lambda_i(t_1)]\xi_1^i + \sum_{i=1}^{n} [\lambda_0 g_{x_2^i} - \lambda_i(t_2)]\xi_2^i + c\int_{t_1}^{t_2} [H(\tau, x_0(\tau), u(\tau), \lambda(\tau)) - H(\tau, x_0(\tau), u_0(\tau), \lambda(\tau))]\, d\tau \geqq 0. \tag{5.4}$$

For $c = 0$ we obtain

$$\sum_{i=1}^{n} [\lambda_0 g_{x_1^i} + \lambda_i(t_1)]\xi_1^i + \sum_{i=1}^{n} [\lambda_0 g_{x_2^i} - \lambda_i(t_2)]\xi_2^i \geqq 0. \tag{5.5}$$

Since B possesses a tangent hyperplane at $\eta(x_0)$, (5.5) holds with equality, and this is relation (P_5) of Pontryagin's necessary conditions when $\tau_1 = \tau_2 = 0$. In particular, (5.4) and (5.5) yield, for $c = 1$:

$$\int_{t_1}^{t_2} [H(\tau, x_0(\tau), u(\tau), \lambda(\tau)) - H(\tau, x_0(\tau), u_0(\tau), \lambda(\tau))]\, d\tau \geqq 0. \tag{5.6}$$

Proof of (P_2). Let $u(t)$ be any bounded measurable function with $u(t) \in U(t)$ a.e. in $[t_1, t_2]$. Let

$$\Delta_u(t) = H(t, x_0(t), u(t), \lambda(t)) - H(t, x_0(t), u_0(t), \lambda(t)).$$

Then $\Delta_u(t)$ is measurable in $[t_1, t_2]$, and for a.a. t in $[t_1, t_2]$,

$$\frac{d}{dt} \int_{t'}^{t} \Delta_u(t)\, dt = \Delta_u(t), \tag{5.7}$$

where $t_1 \leqq t' \leqq t_2, t' \neq t$. Let $\bar{t}$ be such a point. We wish to show that $\Delta_u(\bar{t}) \geqq 0$. To this end, let us choose an arbitrarily small positive h, $t_1 \leqq \bar{t} - h < \bar{t}$, and consider the "mixed control" $u_h(t)$,

$$u_h(t) = \begin{cases} u_0(t), & t \in [t_1, t_2] - [\bar{t} - h, \bar{t}], \\ u(t), & t \in [\bar{t} - h, \bar{t}]. \end{cases}$$

Then $v = (1, u_h, 0)$ is a variation, and hence (5.6) and (5.7) yield

$$0 \leqq \int_{\bar{t}-h}^{\bar{t}} \Delta_{u_h}(t)\, dt = \int_{\bar{t}-h}^{\bar{t}} \Delta_u(t)\, dt = h(\Delta_u(\bar{t})) + o(h).$$

Dividing by $h > 0$, we obtain

$$0 \leqq \Delta_u(\bar{t}) + o(h)/h,$$

and hence, by taking $h \to 0^+$, this yields $\Delta_u(\bar{t}) \geqq 0$. Statement (P_2) is thereby proved. Property (P_2) is of some relevance since no requirement was needed for its proof on the variable closed set $U(t) \subset E_m$ but there are bounded measurable functions u with $u(t) \in U(t)$ a.e. in $[t_1, t_2]$.

Proof of (P_3). Since $U(t) = U$ is a subset of E_m, there is a countable subset U_c of U such that the closure of U_c, cl U_c, is U. Let $U_c = \{u_1, u_2, \cdots, u_k, \cdots\}$. Consider the constant controls $u_i(t) = u_i$, $t_1 \leqq t \leqq t_2$, $i = 1, 2, \cdots$. Then for each i, $u_i(t)$ is a measurable bounded function in $[t_1, t_2]$, with $u_i \in U$. Hence (P_2) applies to each of these controls. In particular, for each i, there exists a set $K_i \subset [t_1, t_2]$, measure of $K_i = m(K_i) = 0$, such that

$$H(t, x_0(t), u_0(t), \lambda(t)) \leqq H(t, x_0(t), u(t), \lambda(t)) \tag{5.8}$$

holds for $u(t) = u_i(t)$ in $[t_1, t_2] - K_i$. Let $K = \cup_i K_i$. Then $m(K) = 0$. Let $G = [t_1, t_2] - K$. We shall now show that (5.8) holds for $t \in G$. Choose any $t_0 \in G$. Since cl $U_c = U$, there exists a (minimizing) subsequence $[u_{k_j}]$ of $[u_k]$ such that

$$H(t_0, x_0(t_0), u_{k_j}, \lambda(t_0)) \to \inf_{u \in U} H(t_0, x_0(t_0), u, \lambda(t_0)) = M(t_0, x_0(t_0), \lambda(t_0)) \tag{5.9}$$

as $j \to \infty$. Moreover, from (5.8),

$$H(t_0, x_0(t_0), u_{k_j}, \lambda(t_0)) = H(t_0, x_0(t_0), u_{k_j}(t_0), \lambda(t_0)) \geqq H(t_0, x_0(t_0), u_0(t_0), \lambda(t_0)), \tag{5.10}$$

$j = 1, 2, \cdots$. Hence (5.9) and (5.10) yield

$$H(t_0, x_0(t_0), u_0(t_0), \lambda(t_0)) \leqq \inf_{u \in U} H(t_0, x_0(t_0), u, \lambda(t_0)). \tag{5.11}$$

Since $u(t_0) \in U$, (5.11) holds with equality. Since t_0 was chosen arbitrarily in G, and $m(G) = t_2 - t_1$, (P_3) is thereby proved.

The proof of (P_4) may be found in [7] or [10], and hence is omitted.

Remark 5.1. The strong form of the Pontryagin necessary conditions can be proved even for time-dependent control spaces $U(t)$ under suitable conditions on $U(t)$. For in the proof of (P_3) we only required that the class Γ of admissible controls be "separable"; that is, (Q) there exists a countable collection $\{u_i\}$ of controls from Γ such that for a.a. t_0 in $[t_1, t_2]$, and any value w in $U(t_0)$, there is at least one subsequence $\{u_{i_k}\}$ such that $u_{i_k}(t_0) \to w$ as $k \to \infty$.

The usual minimum principle then follows by applying the weak minimum principle for each $u_i(t)$, $t_1 \leqq t \leqq t_2$, $i = 1, 2, \cdots$. As just observed, it can be shown that Γ is separable if $U(t) = U$. However, it is clear that whenever property (Q) holds, then the Pontryagin necessary conditions, and its corollaries, hold in the usual form (P_3), as, for example, it is assumed that the "boundaries" of $U(t)$ are continuous functions in $[t_1, t_2]$.

6. Systems with state variable in C. We consider here control systems whose state variables for every t are in C, the set of continuous functions varying over a given set I. In contrast, the admissible controls are taken as measurable functions of t alone. The cost functional is now taken in the form $I[x, u] = g(t_1, \int_I x(t_1, a)\, dP, t_2, \int_I x(t_2, a)\, dP)$, or $\int_I g(t_1, x(t_1, a), t_2, x(t_2, a))\, dP$, where h is a continuous real-valued function and P is a finite measure over I. (See Remark 7.1(a).) The existence of optimal pairs for such systems is discussed in [2], [3]; we show here how the previous proof of Pontryagin's necessary condition may be modified to encompass these systems. To this end, let $I \subset E_1$, I compact. Let $A(a)$, $a \in I$, denote the constraint sets, where $A(a)$ are compact subsets of the tx-space $E_1 \times E_n$, with t in E_1, and $x = (x^1, \cdots, x^n)$, the space variable, in E_n. We assume that $A = \bigcup_{a \in I} A(a)$ is compact. Let $U(t)$, the control set, be a subset of the u-space E_m, $u = (u^1, \cdots, u^m)$, the control variable. Let $M(a) = \{(t, x, u) : (t, x) \in A(a), u \in U(t)\}$ and $M = \bigcup_{a \in I} M(a) = \{(t, x, u) : (t, x) \in A, u \in U(t)\}$ be compact subsets of E_{1+n+m}, and let $f = (f_1, \cdots, f_n)$ be a continuous vector function from M into E_n (see also Remark 7.1(b)). Let the boundary set B be a closed set of points of E_{2n+2}. Let P be a finite measure defined from I into B, and for $z(a) = (z^1(a), \cdots, z^n(a))$, let $R(z(a)) = \int_I z(a)\, dP = (\int_I z^1(a)\, dP, \cdots, \int_I z^n(a)\, dP) \in E_n$.

We shall consider the class Ω of all pairs, $x(t, a), u(t)$, $t_1 \leqq t \leqq t_2$, $a \in I$, called admissible pairs, of the family of vector functions $x(t, a)$, and the vector functions $u(t)$, satisfying the following conditions:

(a) $x(t, a)$ is absolutely continuous in $[t_1, t_2]$ for each $a \in I$;

(b) $x(t, a)$ is continuous in I for each t in $[t_1, t_2]$ (see Remark 6.1);

(c) $u(t)$ is measurable in $[t_1, t_2]$;

(d) $(t, x(t, a)) \in A(a)$, $t_1 \leqq t \leqq t_2$, for each $a \in I$;

(e) $(t_1, Rx(t_1, a), t_2, Rx(t_2, a)) \in B$;

(f) $u(t) \in U(t)$, $t_1 \leqq t \leqq t_2$;

(g) the state equation $d(x(t, a))/dt = f(t, x(t, a), u(t))$ is satisfied a.e. in $[t_1, t_2]$

for each $a \in I$ (see also Remark 7.1(b)). Let $\eta(x) = (t_1, Rx(t_1, a), t_2, Rx(t_2, a))$, $a \in I$. Hence $\eta(x) \in E_{2n+2}$. Let h be a continuous real-valued function defined on B. The functional $I[x, u] = g(\eta(x)) = g(t_1, Rx(t_1, a), t_2, Rx(t_2, a))$ is called the cost functional.

We seek the absolute minimum of $I[x, u]$ in the class Ω. If $(\tilde{x}, \tilde{u}) \in \Omega$ has the property that $I[\tilde{x}, \tilde{u}] \leqq I[x, u]$ for all $(x, u) \in \Omega$, then we say that $\tilde{x}, \tilde{u}$ is an optimal pair, and we say that $\tilde{u}$ is an optimal control, and $\tilde{x}$ is an optimal trajectory. Though the optimal pair $\tilde{x}, \tilde{u}$ may not be unique in Ω, the value of the cost functional $I[\tilde{x}, \tilde{u}]$ is the same for all optimal pairs.

Remark 6.1. If $x(t_1, a)$, $a \in I$, is in C, and f is Lipschitzian in x uniformly in t and u, then it follows that $x(t, a)$, $a \in I$, will be in C for all t in $[t_1, t_2]$. In particular, if $x(t_1, a)$ is a fixed continuous function over I, and f satisfies the assumptions of the necessary conditions, then condition (b) is automatically satisfied. We now state necessary conditions, which are analogous to Pontryagin's conditions, for a pair to be optimal for such systems.

We shall assume that the following convexity condition is satisfied: (R) For a.a. t, given any continuous function $r(a)$ from I into E_n with $(t, r(a)) \in A$ for each $a \in I$, let $Q(t, r(\cdot))$ be the set of all functions from I into E_n of the form $z(a) = f(t, r(a), u)$, $a \in I$, as u describes $U(t)$, and we assume that $Q(t, r(\cdot))$ is convex. In other words, given any two points $u_1, u_2 \in U(t)$ and $0 \leqq \alpha \leqq 1$, we assume that there is some $u \in U(t)$ such that $\alpha f(t, r(a), u_1) + (1 - \alpha) f(t, r(a), u_2) = f(t, r(a), u)$ for all $a \in I$.

For instance, all functions f of the form $f = A(t, x) + B(t, x)\Phi(t, u)$, where A, B are matrices with continuous entries of dimensions $n \times 1$, $n \times p$, $p \times 1$, satisfy condition (R) provided $\Phi(t, U(t))$ is a convex subset of E_p for every t.

THEOREM 2. *A control system as described above is given. In addition let us assume* $f(t, x, u)$ *possesses continuous partial derivatives* $f_t = (f_{it} = \partial f_i/\partial t$, $i = 1, \cdots, n)$, $f_x = (f_{ix^j}, i, j = 1, \cdots, n)$ *in* M. *Also, we assume that property* (R) *holds. (If* $U(t) = U$, *fixed, this assumption may be eliminated; in particular, Remark* (2.1(a)) *applies to these systems.)*

Let $x_0(t, a), u_0(t), t_1 \leqq t \leqq t_2$, *denote an optimal pair for which*:

(i) *There exists a* $\delta > 0$ *such that the graph of* $x_0, [(t, x_0(t, a)), t_1 \leqq t \leqq t_2]$, *is interior to* $A(a)$ *for each* $a \in I$, *by an amount* δ.

(ii) $u_0(t)$ *is bounded in* $[t_1, t_2]$.

(iii) *The endpoint* $\eta(x_0) = (t_1, Rx_0(t_1, a), t_2, Rx_0(t_2, a)) \in E_{2n+2}$ *of the optimal trajectory* x_0 *is a point of* B, *where* B *possesses a tangent hyperplane* B' *of some dimension* k, $0 \leqq k \leqq 2n + 2$, *whose vectors will be denoted by* $h = (\tau_1, \xi_1, \tau_2, \xi_2)$, *with* $\xi_1 = (\xi_1^1, \cdots, \xi_1^n)$, $\xi_2 = (\xi_2^1, \cdots, \xi_2^n)$, *or in differential form,* $h = (dt_1, dy_1, dt_2, dy_2)$, *with* $dy_1 = (dy_1^1, \cdots, dy_1^n)$, $dy_2 = (dy_2^1, \cdots, dy_2^n)$.

(iv) g *possesses a differential* dg *at* $\eta(x_0)$, *say*

$$dg = g_{t_1}\tau_1 + \sum_{i=1}^{n} g_{x_1^i}\xi_1^i + g_{t_2}\tau_2 + \sum_{i=1}^{n} g_{x_2^i}\xi_2^i,$$

or

$$dg = g_{t_1}\, dt_1 + \sum_{i=1}^{n} g_{x_1^i}\, dy_1^i + g_{t_2}\, dt_2 + \sum_{i=1}^{n} g_{x_2^i}\, dy_2^i,$$

where $g_{t_1}, \cdots, g_{x_2^n}$ *denote the partial derivatives of* g *with respect to* $t_1, \cdots, x_2^n$, *all evaluated at* $\eta(x_0)$.

Let the Hamiltonian H *be defined by*

$$H(t, x, u, \lambda) = \lambda f(t, x, u) = \lambda_1 f_1(t, x, u) + \cdots + \lambda_n f_n(t, x, u).$$

Then there exists a family of vector functions

$$\lambda(t, a) = (\lambda_1(t, a), \cdots, \lambda_n(t, a)), \qquad t_1 \leqq t \leqq t_2, \quad a \in I,$$

which we shall call multipliers, such that:

(G_1) *For each* $a \in I$, $\lambda(t, a)$ *is absolutely continuous in* $[t_1, t_2]$, *and,*

$$\frac{d\lambda_i(t, a)}{dt} = -H_{x_i}(t, x_0(t, a), u_0(t), \lambda(t, a)),$$

$i = 1, \cdots, n$, *for a.a.* t *in* $[t_1, t_2]$. *Moreover,* $\lambda_i(t_2, a)$ *is independent of* a, *so that* $\lambda_i(t_2, a) = \lambda_i(t_2)$, $a \in I$, $i = 1, \cdots, n$.

(G_2) (Weak minimum principle). *Given any bounded, measurable function* $u(t)$, $u(t) \in U(t)$ *a.e. in* $[t_1, t_2]$, *then for a.a.* t *in* $[t_1, t_2]$,

$$RH(t, x_0(t, a), u_0(t), \lambda(t, a)) \leqq RH(t, x_0(t, a), u(t), \lambda(t, a)).$$

(G_3) (Usual minimum principle). *Let* $U(t) = U$, $t_1 \leqq t \leqq t_2$, *a fixed closed subset of* E_m. *Then for a.a. fixed* t *in* $[t_1, t_2]$,

$$M(t) = RH(t, x_0(t, x_0(t, a), u_0(t), \lambda(t, a))$$

for a.a. t *in* $[t_1, t_2]$, *with*

$$M(t) = \inf_{u \in U} RH(t, x_0(t, a), u, \lambda(t, a)).$$

(G_4) *The function* $M(t)$ *is absolutely continuous in* $[t_1, t_2]$, *and*

$$\frac{dM(t)}{dt} = RH_t(t, x_0(t, a), u_0(t), \lambda(t, a))$$

for a.a. t *in* $[t_1, t_2]$.

(G_5) (Transversality relation). *There is a constant* $\lambda_0 \geqq 0$ *such that*

$$(\lambda_0 g_{t_1} - M(t_1))\, dt_1 + \sum_{i=1}^{n} (\lambda_0 g_{x_1^i} + R\lambda_i(t_1, a))\, dy_i^1$$

$$+ (\lambda_0 g_{t_2} + M(t_2))\, dt_2 + \sum_{i=1}^{n} (\lambda_0 g_{x_2^i} - R\lambda_i(t_2, a))\, dy_i^2 = 0,$$

for every vector $h = (dt_1, dy_1, dt_2, dy_2)$ *in* B'. *Moreover,* $(\lambda_0, \lambda(t_2, a))$ *is a nonzero vector, not varying with* $a \in I$.

In the next section, we shall indicate how the previous proof of Pontryagin's principle may be altered to accommodate this new system.

7. Outline of proof of Theorem 2. We again assume t_1 and t_2 are fixed. The main difference between this proof, and the previous one, is in the concept of the variation v. In particular, instead of varying the points $x(t_1)$ and $x(t_2)$ of an admissible trajectory, we now vary $x(t_1, a)$, $x(t_2, a)$, $a \in I$, as functions of a alone. More precisely, let $x(t, a)$, $u(t)$, $t_1 \leqq t \leqq t_2$, $a \in I$, be an admissible pair. Let $\theta(x, a) = (t_1, x(t_1, a), t_2, x(t_2, a))$, and let $\eta(x) = R\theta(x, a) = (t_1, Rx(t_1, a), t_2, Rx(t_2, a))$.

Then, by a variation v, we shall mean a triple $v = (c, u, h)$ made up of a non-negative number c, a bounded measurable m-vector function $u(t) = (u^1(t), \cdots, u^m(t))$, $t_1 \leqq t \leqq t_2$, with $u(t) \in U(t)$ for all t in $[t_1, t_2]$, and a $(2n+2)$-vector $h = (\tau_1, \xi_1, \tau_2, \xi_2) \in B'$, that is, a tangent vector to B at the point $\eta(x_0)$. Since t_1, t_2 are fixed, $\tau_1 = \tau_2 = 0$. Again, we can think of $h = (0, \xi_1, 0, \xi_2)$ as the vector tangent to a curve C of class C^1 lying in B and issued from $\eta(x_0)$, say in this case $C: (t_1, X_1(z), t_2, X_2(z))$, $0 \leqq z \leqq 1$, and hence

$$(t_1, X_1(z), t_2, X_2(z)) \in B, \qquad 0 \leqq z \leqq 1, \tag{7.1}$$

$$Rx_0(t_1, a) = X_1(0), \quad Rx_0(t_2, a) = X_2(0), \quad \xi_1 = X_1'(0), \quad \xi_2 = X_2'(0),$$

where $X_1(z)$, $X_2(z)$ are continuously differentiable in $[0, 1]$. In addition, we can think of $h = (0, \xi_1, 0, \xi_2)$ as the tangent to the curves $D(a)$, $a \in I$, of class C^1 lying in E_{2n+2}, at the point $\theta(x_0, a)$ for each $a \in I$, where $D(a)$ is defined as:

$$(t_1, Y_1(z, a), t_2, Y_2(z, a)), \qquad 0 \leqq z \leqq 1,$$

where, for each $a \in I$,

$$Y_i(z, a) = x_0(t_i, a) + X_i(z) - Rx_0(t_i, a), \qquad i = 1, 2, \tag{7.2}$$

and hence

$$Y_1(0, a) = x_0(t_1, a), \quad Y_2(0, a) = x_0(t_2, a), \quad \xi_1 = Y_1'(0, a), \ \xi_2 = Y_2'(0, a),$$

for each $a \in I$.

As before, we extend $X_1(z), X_2(z)$ in the whole interval $[-1, 1]$, so that $X_1(z), X_2(z)$ are still continuously differentiable in the whole interval $[-1, 1]$ (without requesting that the added arc of the curve C lie in B). With this extension we can say that ξ_1 is the tangent to the curve $C': x = X_1(z)$ at $z = 0$, and ξ_2 is the tangent to the curve $C'': x = X_2(z)$ at $z = 0$. This extension induces a similar extension for each curve $D(a), a \in I$, with ξ_1 tangent to each curve $D_1(a): x = Y_1(z, a)$ at $z = 0$, and ξ_2 tangent to the curves $D_2(a): x = Y_2(z, a)$ at $z = 0$, for each $a \in I$.

Again, we extend this construction to an arbitrary system of s variations as before, say $v_\sigma = (c_\sigma, u_\sigma, h_\sigma)$, $\sigma = 1, 2, \cdots, s$. That is, we introduce the vector variable $z = (z_1, \cdots, z_s)$ varying in the interval $J = [0 \leqq z_\sigma \leqq 1, \sigma = 1, 2, \cdots, s]$, and vector functions $X_1(z), X_2(z)$ defined in J, with $Y_1(z, a)$, $Y_2(z, a)$ defined as in (7.2), such that, for each $a \in I$,

$$\begin{aligned}
&(t_1, X_1(z), t_2, X_2(z)) \in B \quad \text{for } z \in J, \\
&x_0(t_1, a) = Y_1(0, a), \qquad Rx_0(t_1, a) = X_1(0), \\
&x_0(t_2, a) = Y_2(0, a), \qquad Rx_0(t_2, a) = X_2(0), \\
&\xi_{1\sigma} = X_{1z_\sigma}(0) = Y_{1z_\sigma}(0, a), \qquad \xi_{2\sigma} = X_{2z_\sigma}(0) = Y_{2z_\sigma}(0, a),
\end{aligned} \tag{7.3}$$

where $Y_1(z, a)$, $Y_2(z, a)$, $X_1(z)$, $X_2(z)$ are continuously differentiable for z in J for each $a \in I$, and where $Y_{1z_\sigma}(0, a)$, $Y_{2z_\sigma}(0, a)$, $X_{1z_\sigma}(0)$, $X_{2z_\sigma}(0)$ denote the partial derivatives of Y_1, Y_2, X_1, X_2 with respect to z_σ at $0 = (0, \cdots, 0)$. Note that $(X_1, X_2)(Y_1, Y_2)$ represent the curves C_σ, D_σ above, respectively, when z describes the interval $0 \leqq z_\sigma \leqq 1$ of the z_σ-axis, $\sigma = 1, 2, \cdots, s$. We extend the functions $Y_1(z, a)$, $Y_2(z, a)$, $X_1(z)$, $X_2(z)$ in the whole interval $V = [-1 \leqq z_\sigma \leqq 1$, $\sigma = 1, 2, \cdots, s]$ so that $Y_1(z, a)$, $Y_2(z, a)$, $X_1(z)$, $X_2(z)$ are still continuously

differentiable in all of V (without requesting that the new parts of the manifold so added for X_1, X_2 lie in B, that is, we require that (7.3) be satisfied for $z \in J$, but not necessarily for z in V).

For every $z \in V$ we consider now the differential system with initial conditions

$$\begin{aligned} &\frac{dx(t,z,a)}{dt} = q_z(t, x(t,z,a)), && t_1 \leqq t \leqq t_2, \\ &x(t_1, z, a) = Y_1(z, a), \end{aligned} \tag{7.4}$$

for each $a \in I$, where $q_z(t, x)$ is a vector function of t and x depending on the parameter $z \in V$ defined by

$$q_z(t,x) = \left(1 - \sum_\sigma c_\sigma z_\sigma\right) f(t, x, u_0(t)) + \sum_\sigma c_\sigma z_\sigma f(t, x, u_\sigma(t)), \tag{7.5}$$

$t_1 \leqq t \leqq t_2$. Note that for $z = 0 = (0, \cdots, 0)$ we have $q_0(t,x) = f(t, x, u_0(t))$, $Y_1(0, a) = x(t_1, 0, a) = x(t_1, a)$. Thus for $z = 0$, $x(t, 0, a)$ satisfies the same differential system and initial conditions as $x_0(t, a)$ for each $a \in I$; and hence, by uniqueness theorems, $x(t, 0, a) = x_0(t, a)$ for all $t_1 \leqq t \leqq t_2$, $a \in I$.

It can now be shown that since the expression (7.5) for $q_z(t, x)$ is linear in $z = (z_1, \cdots, z_s)$, that $Y_1(z, a)$ is continuous in the pair $(z, a) \in V \times I$, and that the graph $[(t, x_0(t, a)),\ t_1 \leqq t \leqq t_2]$ of x_0 has been assumed to lie in the interior of $A(a)$ by an amount of at least δ, $0 < \delta \leqq 1$, there then exists a number γ, $0 < \gamma \leqq 1$, such that $|x(t, z, a) - x(t, 0, a)| = |x(t, z, a) - x_0(t, a))| \leqq \delta/2$ for $|z_\sigma| < \gamma$, $t_1 \leqq t \leqq t_2$, and all $a \in I$. Thus, if we denote by V_γ the interval $V_\gamma = [-\gamma \leqq z_\sigma \leqq \gamma,\ \sigma = 1, \cdots, s]$, we conclude that for every $z \in V_\gamma$, and $a \in I$, $x(t, z, a)$ exists in the whole interval $[t_1, t_2]$, and its graph lies in $A(a)$ for each $a \in I$. For $z \in V_\gamma \cap J$, and again $\gamma > 0$ sufficiently small, we have $c_\sigma z_\sigma \geqq 0$, $\sigma = 1, 2, \cdots, s$, $1 - \sum_\sigma c_\sigma z_\sigma \geqq 0$; and hence $q_z(t, x(t, z, a))$ is a convex combination of the $s + 1$ functions of a, $f(t, x(t, z, a), u_0(t))$, $f(t, x(t, z, a),\ u_\sigma(t))$, $\sigma = 1, 2, \cdots, s$, all in the convex set $Q(t, x(t, z, \cdot))$, and hence $q_z(t, x, \cdot) \in Q(t, x(t, z, \cdot))$, or $dx(t, z, \cdot)/dt \in Q(t, x(t, z, \cdot))$ for almost all t in $[t_1, t_2]$ and $z \in V_\gamma \cap J$. Thus, for a.a. $t \in [t_1, t_2]$ and $z \in V_\gamma \cap J$, there is a $\bar{u}(t, z) \in U(t)$ such that $dx(t, z, \cdot)/dt = f(t, x(t, z, \cdot), \bar{u}(t, z))$. We need only to prove that there is also a function $u(t, z)$, $t_1 \leqq t \leqq t_2$, which has the same property and is measurable in $[t_1, t_2]$. To this purpose let us denote by $\mathfrak{M}$ the set of all triples (t, φ, u) with $t \in [t_1, t_2]$, $\varphi \in C = C(I)$, $(t, \varphi(a)) \in A(a)$ for all $a \in I$, $u \in U(t)$. This set $\mathfrak{M}$ is a separable metric space as a subset of the separable metric space $E_1 \times C \times E_m$ [J. Dieudonné, *Foundations of Modern Analysis*, 1960, (3.10.9)]. Let $\mathfrak{N}$ be the space $E_1 \times C \times C$. Here $\mathfrak{N}$ is a Hausdorff space, and for all $t \in [t_1, t_2]$ a map $K : \mathfrak{M} \to \mathfrak{N}$ is defined by $(t, \varphi, u) \to (t, \varphi, f(t, \varphi, u))$. This map K is continuous in the topologies of $\mathfrak{M}$ and $\mathfrak{N}$. For every $z \in V_\gamma \cap J$ let us consider the measurable map $Y_z : [t_1, t_2] \to \mathfrak{N}$ defined by $t \to (t, x(t, z, \cdot), x'(t, z, \cdot))$. We have just proved that for almost all $t \in [t_1, t_2]$, $Y_z(t) \in K(\mathfrak{M}) \subset \mathfrak{N}$. By E. J. McShane's and R. B. Warfield's implicit function theorem 4 of [9] we know that there exists a measurable function $X_z : [t_1, t_2] \to \mathfrak{M}$ such that $Y_z = KX_z$. In other words, for each $z \in V_\gamma \cap I$ there is a measurable vector function $u(t, z)$, $t_1 \leqq t \leqq t_2$, such that $u(t, z) \in U(t)$ and

$$dx(t, z, a)/dt = f(t, x(t, z, a), u(t, z)) \tag{7.6}$$

a.e. in $[t_1, t_2]$, for each $a \in I$.

Thus, as concluded in § 3, $x(t, z, a), u(t, z)$ satisfies all the conditions of an admissible pair, except possibly the terminal boundary conditions. Moreover, for $z \in V$, $z \geqq 0$, if $Rx(t_2, z, a) = X_2(z)$, then $\eta(x) \in B$ and thus $(x(t, z, a), u(t, z)) \in \Omega$. We also obtain the natural analogue of the variational equations. For we can again formally differentiate (7.4) with respect to z_σ for each $a \in I$. Moreover, $\partial x^i(t, z, a)/\partial z_\sigma$, $i = 1, \cdots, n, \sigma = 1, \cdots, s$, are continuous functions of $a \in I$ for each fixed (t, z) in $[t_1, t_2] \times V$. Hence, by the Lebesgue dominated convergence theorem,

$$R\frac{\partial x^i(t, z, a)}{\partial z_\sigma} = \frac{\partial Rx^i(t, z, a)}{\partial z_\sigma}, \tag{7.7}$$

$\sigma = 1, \cdots, s$, for $t \in [t_1, t_2]$. Let

$$W_\sigma(t, a) = \left(\frac{\partial x(t, z, a)}{\partial z_\sigma}\right)_{z=0}$$

for each $(t, a) \in [t_1, t_2] \times I, \sigma = 1, \cdots, s$. Then we conclude from the above observations that $W_\sigma(t, a)$ satisfies the same variational equation as $\partial x^i(t, z)/\partial z_\sigma$ of § 3; that is, we obtain

$$\begin{aligned} \frac{dW_\sigma(t, a)}{dt} &= -c_\sigma f(t, x_0(t, a), u_0(t)) + c_\sigma f(t, x_0(t, a), u_\sigma(t)) \\ &\quad + f_x(t, x_0(t, a), u_0(t))W_\sigma(t, a), \qquad t_1 \leqq t \leqq t_2, \\ W_\sigma(t_1, a) &= \xi_{1\sigma}, \qquad \sigma = 1, \cdots, s. \end{aligned} \tag{7.8}$$

We now form a convex cone K analogous to that of § 4. An important property of this cone is that, as in § 4, the cone is made up of points in E_{n+1}, and hence we can again consider support planes to K in E_{n+1}.

For every variation $v = (c, u, h)$ with $h = (0, \xi_1, 0, \xi_2)$, let us consider the $(n + 1)$-vector $\tilde{G}(v) = (G^0(v), G(v)) = (G^0(v), G^1(v), \cdots, G^n(v))$ (in the $z_0 z_1 \cdots z_n$ space E_{n+1}) defined by

$$G^0(v) = g_{x_1}\xi_1 + g_{x_2}\xi_2, \qquad G(v) = w(t_2, v) - \xi_2,$$

where g_{x_1}, g_{x_2} are evaluated at the point $\eta(x_0)$, and $w(t_2, v) = RW(t, a, v)$, where $W(t, a, v)$ denotes the solution of the differential system and initial data (7.8) with $v_\sigma = v = (c, u, h)$, $h = (0, \xi_1, 0, \xi_2)$. We shall denote by $K \subset E_{n+1}$ the set of all such vectors $G(v)$ in E_{n+1}.

By considering the variational system (7.8) for each $a \in I$, and by using the same methods of the proof of Lemma 4.1, we can prove that K is a convex cone with vertex at $(0, \cdots, 0)$. Again, the central result is that the vector $\zeta = (-1, 0, \cdots, 0)$ is not interior to K, and once more, this follows in the same way as Lemma 4.2, using (7.7). For again, assume that $(-1, 0, \cdots, 0)$ is interior to K. Then for some $\delta > 0$ sufficiently small, the $n + 1$ points in E_{n+1}:

$$\begin{gathered} (-1, -\delta, 0, \cdots, 0), (-1, 0, -\delta, 0, \cdots, 0), \cdots, (-1, 0, \cdots, 0, -\delta), \\ (-1, \delta, \cdots, \delta) \end{gathered} \tag{7.9}$$

belong to K, and hence there are variations $v_1, v_2, \cdots, v_{n+1}$ such that the corresponding vectors $\tilde{G}(v_1), \cdots, \tilde{G}(v_{n+1})$ are exactly the vectors (7.9) with $v_\sigma = (c_\sigma, u_\sigma, h_\sigma)$,

$h_\sigma = (0, \xi_{1\sigma}, 0, \xi_{2\sigma})$, $\sigma = 1, \cdots, n+1$. We then take $s = n+1$ at the beginning of this proof and denote by $Y_1(z,a)$, $Y_2(z,a)$, $X_1(z)$, $X_2(z)$, $z = (z_1, \cdots, z_{n+1}) \in V_\gamma$, the corresponding functions, and by $x(t,z,a)$ the corresponding solution of equation (7.4) with initial values $x(t_1, z, a) = Y_1(z,a)$. As before we have to compare the expectation of the end values of $x(t,z,a)$, or $x(t_2,z,a)$, with $X_2(z)$, and the value of the functional $g(\eta(x(t,z,a)))$ with $g(\eta(x_0)) = g_{\min}$. Namely, we consider the equations

$$\begin{aligned} &g(t_1, X_1(z), t_2, X_2(z)) + z_0 - g_{\min} = 0, \\ &r_i(t_2, z) - X_{2i}(z) = 0, \end{aligned} \tag{7.10}$$

$i = 1, \cdots, n$, where $r_i(t_2, z) = Rx_i(t_2, z, a)$. These $n+1$ equations (in the $n+2$ unknowns $z_1, \cdots, z_{n+1}, z_0$) are obviously satisfied by $z_1 = \cdots = z_{n+1} = z_0 = 0$ since then $x(t_1, 0, a) = x_0(t_1, a) = Y_1(0,a)$, $x(t_2, 0, a) = x_0(t_2, a) = Y_2(0,a)$, and thus $Rx(t_1, 0, a) = RY_1(0,a) = X_1(0)$, $Rx(t_2, 0, a) = RY_2(0,a) = X_2(0)$; also, $g(\eta(x_0)) = g_{\min}$. At the point $(0, 0, \cdots, 0)$ the partial derivatives of the first members with respect to z_σ are respectively

$$\sum_{j=1}^{n} (g_{x_1^j} X^j_{1z_\sigma} + g_{x_2^j} X^j_{2z_\sigma})_{z=0} = \sum_{j=1}^{n} (g_{x_1^j} \xi^j_{1\sigma} + g_{x_2^j} \xi^j_{2\sigma}) = G^0(v_\sigma),$$

and by (7.7),

$$R\left[\frac{\partial x^i(t_2, z, a)}{\partial z_\sigma}\right]_{z=0} - [X^i_{2z_\sigma}(z)]_{z=0} = RW^i(t_2, a, v_\sigma) - \xi^i_{2\sigma} = G^i(v_\sigma),$$

$i = 1, \cdots, n$; $\sigma = 1, \cdots, n+1$. In other words, the $(n+1) \times (n+1)$ functional determinant of the $n+1$ equations (7.10) with respect to the $n+1$ variables $z_1, z_2, \cdots, z_{n+1}$ is the determinant of the $n+1$ vectors (7.9) and this determinant is $(-1)^{n+1}(n+1)\delta^n \neq 0$.

Using the implicit function theorem of calculus, we can show, as we did in § 4, that there are variations satisfying the appropriate boundary conditions, for which the cost functional is lower than $\min g$. Since these variations yield admissible pairs, this leads to a contradiction. Hence $\zeta \notin \operatorname{int} K$, and we conclude as before that there are numbers $\chi_0, \chi_1, \cdots, \chi_n$, not all zero, $\chi_0 \geqq 0$, such that $\sum_{i=0}^{n} \chi_i G^i(v) \geqq 0$ for all variations v. As noted previously, the necessary conditions now follow by taking a supporting hyperplane to K. In particular, the analogue of (5.1) for these systems is

$$\frac{dW(t,a)}{dt} = A(t,a)W(t,a) + b(t,a), \qquad W(t_1, a) = \xi_1, \tag{7.11}$$

where $A(t,a)$ is the $n \times n$ matrix $A(t,a) = f_x(t, x_0(t,a), u_0(t)) = (f_{ix^j}, i, j = 1, \cdots, n)$, and b is the n-vector

$$b(t,a) = -cf(t, x_0(t,a), u_0(t)) + cf(t, x_0(t,a), u(t)).$$

We denote by χ the constant n-vector $\chi = (\chi_1, \cdots, \chi_n)$, and by $\Phi(t,a) = (\phi_{ij}(t,a), i,j = 1, \cdots, n)$, $t_1 \leqq t \leqq t_2$, any fundamental system of solutions of the homogeneous linear system $dW/dt = A(t,a)W$, $t_1 \leqq t \leqq t_2$. Let $\Phi^{-1}(t,a) = \Psi(t,a) = (\psi_{ij}(t,a), i, j = 1, \cdots, n)$, the inverse matrix of Φ. Again D_{-1} is the transpose of matrix D.

With these notations the solution $W(t, a, v)$, $t_1 \leqq t \leqq t_2$, of the variational equation and initial data (7.11) can be written in the explicit form, for each $a \in I$,

$$W(t,a,v) = \Phi(t,a)[\Phi^{-1}(t_1,a)\xi_1 + \int_{t_1}^{t} \Phi^{-1}(t,a)[-cf(\tau, x_0(\tau,a), u_0(\tau)) + cf(\tau, x_0(\tau,a), u(\tau))]\,d\tau], \tag{7.12}$$

that is, (7.12) is the analogue of (5.2). We define the n-vector $\lambda(t,a) = (\lambda_1(t,a), \cdots, \lambda_n(t,a))$ of the multipliers by taking, for each $a \in I$, $\lambda_0(t,a) = \lambda_0 = \chi_0$,

$$\lambda(t,a) = (\chi_{-1}\Phi(t_2,a)\Phi^{-1}(t,a))_{-1},$$

$t_1 \leqq t \leqq t_2$. In particular, we have immediately $\lambda(t_2,a) = \chi$ for all $a \in I$, and $\lambda(t,a)$ is absolutely continuous in $[t_1, t_2]$ for each $a \in I$.

All the statements of Theorem 2 now follow by using the arguments of § 5. In particular, (5.4) now becomes

$$\sum_{i=1}^{n} [\lambda_0 g_{x_1^i} + R\lambda_i(t_1,a)]\xi_1^i + \sum_{i=1}^{n} [\lambda_0 g_{x_2^i} - R\lambda_2(t_2,a)]\xi_2^i + c\int_{t_1}^{t_2} [RH(t, x_0(t,a), u(t), \lambda(t,a)) - RH(t, x_0(t,a), u_0(t), \lambda(t,a))]\,dt \geqq 0, \tag{7.13}$$

and as in § 5, Pontryagin's necessary condition now follows. This completes the proof of Theorem 2. Clearly, the remarks of § 5 again pertain to these systems.

Remark 7.1. (a) The above analysis can be extended to systems with fixed initial conditions for which the cost functional is of the form $I = Rg(t_2, x(t_2,a)) = \int_I g[t_2, x(t_2,a)]\,dP$. In particular, (G_1)–(G_4) remain unchanged, and (G_5) becomes (G_5'):

$$(\lambda_0 g_{t_2} + M(t_2))\,dt_2 + \sum_{i=1}^{n} (\lambda_0 g_{x_2^i} - \lambda_i(t_2,a))\,dy_2^i = 0$$

for every vector $h = (dt_2, dy_2)$ tangent to the terminal boundary set $S \subset E_{1+n}$ at $(t_2, Rx_0(t_2,a))$, where the partial derivatives of g are evaluated at $(t_2, x_0(t_2,a))$ for each $a \in I$ (see [2, Chap. 8]).

(b) The control systems examined in this section also include systems in which the state function f contains a "perturbation" $\eta(t,b) \in E_s$, that is, whose state equation is given by

$$dx(t,a,b)/dt = f(t, x(t,a,b), u(t), \eta(t,b)),$$

where η satisfies smoothness conditions similar to those of x. The operator R is now defined on the class of continuous functions of (a,b) from $I_1 \times I_2$ into E_n, and for such systems, Theorem 2 again holds except that x and λ are now functions of (a,b) for each instant t. See [2, Chap. 8] and [3] for details.

8. Stochastic systems. Let us now consider a class of control systems with stochastic boundary conditions and stochastic state equations. Specifically, let us particularize the control systems considered thus far by setting the initial state $x(t_1,a)$ equal to a fixed continuous function $i(a)$ over I, with t_1 fixed, and by writing

the cost functional as an expectation. In particular, let P now denote a probability measure over the Borel sets restricted to I. The initial conditions are t_1 fixed, with $x(t_1, a) = i(a)$, and the terminal boundary conditions are taken as $(t_2, Ex(t_2, a)) \in S$, S the terminal set, which is a closed subset of E_{1+n}, and where the letter E denotes expectation taken with respect to the probability measure P. That is, we ask that the "average value" of the endpoints lie in a prescribed set S. The cost functional is now given by

$$I[x, u] = Eg(t_2, x(t_2, a)) = \int_I g(t_2, x(t_2, a))\, dP,$$

where g is a real-valued function on S. The operator E clearly satisfies the conditions on R, and hence these systems are subsumed by those of § 6 (with Remark 7.1(a)). For these systems, the state equation is a differential equation with stochastic initial conditions $i(a)$. As noted in Remark 7.1(b), these state equations can be generalized to the form

$$dx(t, a, b)/dt = f(t, x(t, a, b), \eta(t, b), u(t)),$$

where now $\eta(t, b)$ is a stochastic process whose sample paths are of class C^1 in t. Again, Remark 2.1(a) applies to these systems.

The central feature of these systems is that the controls $u(t)$ are functions of t alone, while the resulting trajectories $x(t, a)$, or $x(t, a, b)$, vary with t, and stochastically with a and b. Such systems may arise if we are in a "period of ignorance" and are unable to ascertain either the initial conditions or the state of the system during the operating period. Alternately, it may be necessary, due to extensive setup procedures, to determine $u(t)$, for all t, before the initial conditions are known, and where, due to the lack of technology, we are unable to implement the control in feedback form. (See [2, Chap. 2] and [4] for more extensive discussions of such systems.)

We now consider an example of such systems.

Example. Let us consider the control system with state equation $\dot{x}^1(t, a) = x^2(t, a)$, $\dot{x}^2(t, a) = bu(t)$, and initial conditions $x^1(0, a, b) = 0$, $x^2(0, a, b) = a$, $0 \leqq t \leqq T$, T free. Here a and b are independent random variables with probability densities one, $0 \leqq a \leqq 1$, $1 \leqq b \leqq 2$. We require $|u| \leqq 1$, and we seek admissible pairs $x(t, a, b)$, $u(t)$, $0 \leqq t \leqq T$, which minimize the expectation of $[x^2(T, a, b)]^2$. All the assumptions of Theorem 2 and Remark 7.1 (a) are satisfied, except that we lack compact constraint sets $A(a)$ and A. However, given any admissible trajectory $x(t, a, b)$, $0 \leqq t \leqq T$, it follows from the theory of differential equations that $x(t, a, b)$ is continuous on $[0, T] \times [0, 1] \times [1, 2]$, and hence is bounded there. Thus, given any extremal pair, that is, any pair satisfying the necessary conditions of Theorem 2, there is some compact constraint set A containing this trajectory. Moreover the set $M = A \times [-1, 1]$ is then compact. Hence, the hypotheses of Theorem 2 are satisfied with respect to the set A, and consequently, we may apply Theorem 2. The Hamiltonian H is given by $H = \lambda_1 x^2 + \lambda_2 bu$. Hence the minimum of EH, as a function of u, occurs for $u = -\operatorname{sgn} E(b\lambda_2)$. From (G_1), we obtain $\lambda_1 = 0, \lambda_2 = -\lambda_1$. From the transversality relations, $\lambda_1(T, a) = 0$, $\lambda_2(T, a) = 2x^2(T, a, b)$ (where $\lambda_0 \neq 0$ and hence may be taken as one). Thus, $\lambda_1 \equiv 0$, $\lambda_2 = 2x^2(T, a, b) = \pi_2$. Consequently, $u = -\operatorname{sgn} E(\pi_2 b)$. From (G_3) and (G_4), $M(t) = 0$, $0 \leqq t \leqq T$.

Hence $M(t) = -|E(\pi_2 b)| = 0$ which implies $E(\pi_2 b) = 0$ or

$$E[bx^2(T, a, b)] = 0. \tag{8.1}$$

That is, all extremal pairs must satisfy (8.1). From the state equations, this implies $E[b(a + b\int_0^T u(t)\,dt)] = 0$, from which we obtain

$$\int_0^T u(t)\,dt = -9(28)^{-1}. \tag{8.2}$$

Consequently, any admissible trajectory satisfying (8.1) or (8.2) satisfies $x^2(T, a, b) = -9b(28)^{-1} + a$, and thus attains a cost of $31(336)^{-1}$. For example, we may take

$$u(t) = -1, \quad x^1(t, a, b) = -(2)^{-1}bt^2 + at, \quad x^2(t, a, b) = -bt + a, \quad 0 \leqq t \leqq 9(28)^{-1}. \tag{8.3}$$

This system can be shown to possess at least one solution; see [3] and [4]. It thus follows that all the admissible pairs satisfying (8.1) or (8.2) are optimal.

If T is fixed, and less than $9(28)^{-1}$, then $Ex^2(T, a, b)$ must be positive, and hence $E(\pi_2 b) > 0$. Since $u = -\operatorname{sgn} E(\pi_2 b)$, the optimal pair is given uniquely by (8.3), with $0 \leqq t \leqq T$. If $T \geqq 9(28)^{-1}$, then any admissible pair satisfying (8.1) or (8.2) can easily be shown to be optimal.

Note that we do not obtain these results if we replace a and b by their expected values, $(2)^{-1}$ and $3(2)^{-1}$. For if T is free, we obtain the system $\dot{x}^1 = x^2$, $\dot{x}^2 = 3(2)^{-1}u$, $x^1(0) = 0$, $x^2(0) = (2)^{-1}$, $|u| \leqq 1$, $0 \leqq t \leqq T$, where we wish to minimize $(x^2(T))^2$. Clearly, any admissible pair satisfying $x^2(T) = 0$ is optimal. In terms of our first system, this gives $Ex^2(T, a, b) = (2)^{-1} + 3(2)^{-1}\int_0^T u(t)\,dt = x^2(T) = 0$, which is not (8.1). In particular, the optimal cost is now zero, and the optimal time is $(3)^{-1}$.

REFERENCES

[1] T. S. ANGELL, *Existence theorems for optimal control problems involving functional differential equations*, J. Optimization Theory Appl., 7 (1971), pp. 149–169.

[2] R. F. BAUM, *Optimal control systems with stochastic boundary conditions*, Doctoral thesis, University of Michigan, Ann Arbor, 1969.

[3] ———, *An existence theorem for optimal control systems with state variable in C, and stochastic control problems*, J. Optimization Theory Appl., 5 (1970), pp. 335–346.

[4] ———, *Optimal control systems with stochastic boundary conditions and state equations*, J. Operations Res. Soc. Amer., submitted.

[5] L. CESARI, *Problems of optimization*, to appear.

[6] J. K. HALE, *Ordinary Differential Equations*, Interscience, New York, 1969.

[7] E. J. MCSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.

[8] ———, *Integration*, Princeton University Press, Princeton, N.J., 1944.

[9] E. J. MCSHANE AND R. B. WARFIELD, JR., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.

[10] L. S. PONTRYAGIN, V. G. BOLTJANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

ON CONVERSES TO THE STABILITY THEOREMS FOR DIFFERENCE EQUATIONS*

S. P. GORDON†

Abstract. In this paper, converses to a number of the Lyapunov-type stability theorems for difference equations are obtained. The basic vector difference equation considered is

$$X(n+1) = f(n, X(n)) \tag{1}$$

subject to the initial condition $X(n_0) = x_0$.

It is first shown that if the equilibrium $X = 0$ of the difference equation (1) is stable and if all solutions through any point (n, X) in the domain considered can be uniquely extended back to an initial value at time n_0, then there exists a positive definite discrete Lyapunov function $V(n, X)$ whose total difference

$$\Delta V(n, X) = V(n+1, f(n, X(n))) - V(n, X(n))$$

is negative semidefinite along the discrete trajectories represented by the solutions.

Moreover, if the equilibrium $X = 0$ of the linear difference equation $X(n+1) = A(n)X(n)$, where $A(n)$ is a nonsingular matrix for all n, is asymptotically stable, then it is proved that there exists a positive definite Lyapunov function $V(n, X)$ whose total difference is negative definite along trajectories.

Finally, if the equilibrium $X = 0$ of the difference equation (1) is uniformly asymptotically stable, then there exists a positive definite, decrescent, locally Lipschitzian Lyapunov function $V(n, X)$ whose total difference is negative definite along trajectories.

Introduction. In recent years, considerable attention has been paid to the development of a stability theory for difference equations to parallel that for differential equations. Notable among this research is that of Hahn [1], [2], Halanay [3], Hurt [4] and Kalman and Bertram [5]. However, except for several results given by Halanay [3], the possibility of proving converse results for the various stability theorems has been ignored. The present paper introduces several such results.

Basic concepts and definitions. The difference equation we shall consider is

$$X(n+1) = f(n, X(n)), \tag{1}$$

where $X(n)$ and $f(n, X)$ are $q \times 1$ column vectors and f is a function assuming values in W^q, an arbitrary q-dimensional vector space, and defined on

$$D_{n_0R} = \{(n, X) \in I \times W^q : n \geqq n_0 \geqq 0, 0 \leqq \|X\| \leqq R\}.$$

Here, $\|X\|$ denotes any q-dimensional norm of the vector X. The difference equation will be subject to the initial condition

$$X(n_0) = x_0,$$

and the unique solution of this problem will be denoted by $F(n, n_0, x_0)$.

Moreover, we impose the condition

$$f(n, 0) = 0$$

for all $n \geqq n_0$. Hence, $X(n) = 0$ is the equilibrium for (1).

* Received by the editors February 23, 1970, and in revised form October 18, 1970.

† Department of Mathematics, Queens College of the City University of New York, Flushing, New York 11367. This paper is adapted from a portion of the author's doctoral dissertation submitted to the Faculty of Graduate Studies and Research, McGill University.

We now define the various types of behavior for the solutions of (1) which will be of interest to us in the sequel.

DEFINITION 1. The equilibrium $X = 0$ of (1) is *stable* if, for any $\varepsilon > 0$ and $n_0 \in 1$, there exists a $\delta(\varepsilon, n_0) > 0$ such that $\|x_0\| < \delta$ implies

$$\|F(n, n_0, x_0)\| < \varepsilon$$

for all $n \geqq n_0$.

DEFINITION 2. The equilibrium $X = 0$ of (1) is *asymptotically stable* if it is stable and if, for any $n_0 \in I$, there exists a $\delta_0(n_0) > 0$ such that $\|x_0\| < \delta_0$ implies

$$F(n, n_0, x_0) \to 0$$

as $n \to \infty$.

DEFINITION 3. The equilibrium $X = 0$ of (1) is *uniformly-asymptotically stable* if it is stable and if there is a $\delta_0 > 0$ such that $n_0 \in I$, $\|x_0\| < \delta_0$ imply

$$F(n, n_0, x_0) \to 0$$

uniformly for $n_0 \in I$, $\|x_0\| < \delta_0$ as $n \to \infty$.

As is well known, each of these properties, as well as many other refinements of stability, can be characterized by demonstrating the existence of certain real scalar functions, the so-called Lyapunov functions. In the following, we prove that each of the above three stability properties itself implies the existence of a Lyapunov function with the appropriate properties and with the appropriate conditions on the total difference

$$\Delta V(n, X(n)) = V(n + 1, f(n, X(n))) - V(n, X(n)).$$

We note that this difference is a measure of the growth or decay of the function $V(n, X)$ with regard to increasing n along the discrete trajectories represented by the solutions of (1).

Principal results. We first present the converse to the basic theorem on stability for difference equations. Its proof depends on the fact that all solutions start at some initial time n_0. Moreover, we consider the set D which consists of those points in D_{n_0R} which are specifically determined by the given difference equation. To illustrate this, consider the scalar difference equation $X(n + 1) = \frac{1}{2}X(n)$. The point $(n_0 + 1, \frac{3}{4}R)$ is not in D since it is not the image under the equation of any point in D_{n_0R}.

THEOREM 1. *If the equilibrium $X = 0$ of the difference equation (1) is stable on D_{n_0R} and if the solution of (1) through any point (n, X) in D can be uniquely extended back to time n_0, then there exists a real scalar function $V(n, X)$ for which, on D,*

(a) *$V(n, X)$ is positive definite,*

(b) *$\Delta V(n, X)$ is negative semidefinite.*

Proof. Let (n, X) denote any parameter point in D and let N be the independent variable. Thus, $F(N, n, X)$ represents that solution of the difference equation evaluated at time N which passes through the point (n, X). In order to consider values of N for which $n_0 \leqq N < n$, it is necessary to interpret X as $X = F(n, n_0, x_0)$ for that particular x_0 from which a solution emanates which passes through (n, X). By the hypothesis, this x_0 is unique.

We now consider the scalar function

$$V(n, X) = \|F(n_0, n, X)\|.$$

Since the equilibrium is stable, for all $\varepsilon > 0$, there exists a $\delta > 0$ such that $\|F(n_0, n, X)\| < \delta$ implies that $\|X\| < \varepsilon$. Correspondingly, it follows that for $\|X\| \geqq \varepsilon$, $V(n, X) \geqq \delta > 0$, so that $V(n, X)$ is positive definite.

Furthermore,

$$\Delta V(n, X) = \|F(n_0, n+1, X(n+1))\| - \|F(n_0, n, X(n))\| = 0,$$

since $(n, X(n))$ and $(n+1, X(n+1))$ are two successive points along the same trajectory. As a consequence, it follows that $V(n, X)$ is negative semidefinite and the proof is complete.

Before continuing to a partial converse for the basic theorem on asymptotic stability, we state the following lemma due to Massera [6].

LEMMA (Massera). *Given any real scalar function $g(r)$ defined and positive on every compact interval $J \subset [0, \infty)$ such that $g(r) \to 0$ as $r \to \infty$; and given any real scalar function $h(r)$, defined and continuous, positive and nondecreasing on $[0, \infty)$; then there exists, for any integer $k \geqq 0$, a positive real scalar function $G(r)$, of class C^k, which is increasing together with its first k derivatives on $[0, \infty)$ and with $G^{(i)}(0) = 0$, $i = 0, 1, \cdots, k$, such that, for any real scalar function $g^*(r)$,*

$$0 \leqq g^*(r) \leqq cg(r),$$

for some constant $c > 0$ on $[0, \infty)$, the integrals

$$\int_0^\infty [G^{(i)}(g^*(r))h(r)]\, dr, \qquad 0 \leqq i \leqq k,$$

converge uniformly in g^*.

The analogue of this lemma for difference equations would guarantee the existence of the same scalar function $G(r)$ and the uniform convergence of

$$\sum_{j=0}^{\infty} [G^{(i)}(g^*(j))h(j)], \qquad 0 \leqq i \leqq k.$$

However, the convergence of these sums in the discrete case follows immediately from the convergence of the corresponding integrals, as given in the lemma, by the integral test for the convergence of a series. Hence, it follows that the lemma is valid for the discrete cases we are considering.

THEOREM 2. *If the equilibrium $X = 0$ of the linear difference equation*

$$X(n+1) = A(n)X(n)$$

is asymptotically stable, where the matrix $A(n)$ is nonsingular for all $n \geqq n_0$, then there exists a real scalar function $V(n, X)$ for which, on $D_{n_0 R}$,

(a) *$V(n, X)$ is positive definite,*

(b) *$\Delta V(n, X)$ is negative definite.*

Proof. Denote by $Z(n)$ the fundamental matrix solution of the linear difference equation which satisfies the condition

$$Z(0) = I,$$

the identity matrix. The general solution of the difference equation is then given by

$$F(n, n_0, x_0) = Z(n)Z^{-1}(n_0)x_0.$$

Thus, for $n \geqq n_0$, replacing x_0, n_0, and n respectively by X, n, and N, we find

$$\begin{aligned} \|X\| &= \|Z(n)Z^{-1}(N)F(N, n, X)\| \\ &\leqq \|Z(n)Z^{-1}(N)\| \, \|F(N, n, X)\| . \end{aligned}$$

Now let

$$g(n) = \|Z(n)Z^{-1}(N)\| .$$

For fixed X, $g(n)$ goes to zero as n approaches infinity, since the equilibrium is asymptotically stable.

We now define

$$\begin{aligned} V(n, X) = &\sum_{k=n}^{\infty} G[\|Z(k)Z^{-1}(N)\| \, \|F(N, n, X)\|] \\ &+ \sum_{k=N}^{\infty} G[\|Z(k)Z^{-1}(N)\| \, \|F(N, n, X)\|], \end{aligned}$$

using the discrete form of Massera's lemma. This function is positive definite, since

$$\begin{aligned} V(n, X) &\geqq G[\|Z(n)Z^{-1}(N)\| \, \|F(n, n, X)\|] \\ &\geqq G(\|X\|). \end{aligned}$$

Moreover,

$$\begin{aligned} \Delta V(n, X) = &\sum_{k=n+1}^{\infty} G[\|Z(k)Z^{-1}(N)\| \, \|F(N, n+1, X(n+1))\|] \\ &- \sum_{k=n}^{\infty} G[\|Z(k)Z^{-1}(N)\| \, \|F(N, n, X(n))\| \\ = &- G[\|Z(n)Z^{-1}(N)\| \, \|F(N, n, X)\|] \\ \leqq &- G(\|X\|), \end{aligned}$$

so that $\Delta V(n, X)$ is negative definite, and the theorem is proved.

Finally, we present a converse for the principal theorem on the uniform asymptotic stability of the equilibrium for the difference equation (1).

THEOREM 3. *If the equilibrium $X = 0$ of the difference equation* (1) *is uniformly-asymptotically stable, then there exists a real scalar function $V(n, X)$ which satisfies the following on $D_{n_0 r}$, for some $r \leqq R$:*

(a) *$V(n, X)$ is positive definite;*

(b) *$V(n, X)$ is decrescent;*

(c) *$V(n, X)$ is locally Lipschitzian;*

(d) *$\Delta V(n, X)$ is negative definite.*

Proof. Choose r^*, $0 < r^* < R$, so that for all ε, $0 \leqq \varepsilon \leqq r^*$, there exists a $\delta(\varepsilon) > 0$ such that for $n \in I$ and X with $\|X\| < \delta$,

$$\|F(n + k, n, X)\| < \varepsilon$$

for all $k \geqq 0$. By the uniform-asymptotic stability, there exists a $\delta_0 > 0$ and, for all $\eta > 0$, there exists an integer $v(\eta) \geqq n_0$, such that, for $n \in I$ and $\|X\| < \delta_0$,

$$\|F(n + k, n, X)\| < \eta$$

for $k \geqq v$. Let

$$r = \min(\delta_0, \delta(r^*))$$

and consider the region $D_{n_0 r} \in D_{n_0 R}$ defined by $\{X : \|X\| < r\}$.

Now, given any nonincreasing sequence $\{c_j\}$, $0 < c_j < r$, there exists an increasing divergent sequence of integers $\{n_j\}, n_j(c_j) > 0$, such that $(n, X) \in D_{n_0 r}$ implies that

$$\|F(n + k, n, X)\| < c_j$$

for all $k \geqq n_j$.

Let $g(k)$ be a real scalar function, positive and nonincreasing for $k \geqq 0$, and such that $g(k) \to 0$ as $k \to \infty$, and for all $(n, X) \in D_{n_0 r}$,

$$\|F(n + k, n, X)\| \leqq g(k)$$

on the interval $[0, n_j]$, and let

$$g(n_{j+1}) = c_j$$

for all j. As a result,

$$g(n_{j+1}) \leqq g(k) \leqq g(n_j)$$

for all k in the interval $[n_j, n_{j+1}]$, which implies

$$\|F(n + k, n, X)\| < c_j \leqq g(k)$$

on this interval. This in turn implies

$$\|F(n + k, n, X)\| \leqq g(k)$$

for all $k \geqq 0$.

Now let $G(k)$ be the function associated with $g(k)$, as given in Massera's lemma, where we take $h(k) = 1$. We define the real scalar function

$$V(n, X) = \sum_{k=0}^{\infty} G(\|F(n + k, n, X)\|).$$

This function is well-defined on $D_{n_0 r}$, and by the lemma, $G(s)$ is continuously differentiable, which implies that $V(n, X)$ is also continuously differentiable with respect to X. Also, by the lemma, $\sum_{k=0}^{\infty} G'(\|F(n + k, n, X)\|)$ converges uniformly, and hence is bounded on $D_{n_0 r}$. As a consequence, the vector Φ of partial derivatives of $V(n, X)$ with respect to the components of X is also bounded. Thus, applying a generalized form of the mean value theorem to $V(n, X)$, we obtain

$$\begin{aligned} |V(n, X_1) - V(n, X_2)| &= \|\Phi(n, X^*)\| \, \|X_1 - X_2\| \\ &\leqq M \|X_1 - X_2\|, \end{aligned}$$

where X^* is some value of X between X_1 and X_2, for each n. Thus, $V(n, X)$ is locally Lipschitzian.

Moreover, choosing $X_2 = 0$, we see that

$$|V(n, X_1)| \leqq M\|X_1\|,$$

for every X_1 with $\|X_1\| < r$, so that $V(n, X)$ is also decrescent.

In addition,

$$V(n, X) \geqq G(\|F(n, n, X)\|) = G(\|X\|),$$

so that $V(n, X)$ is positive definite.

Finally, we consider the total difference of $V(n, X)$,

$$\Delta V(n, X) = -G(\|F(n, n, X(n))\|) = -G(\|X\|),$$

so that $\Delta V(n, X)$ is negative definite, and the theorem is proved.

The previous Theorem 3 has been proved by Halanay [3] in the following far more restrictive form.

THEOREM (Halanay). *If there exists a positive monotone increasing function* $m(r)$ *with* $m(0) = 0$ *such that the function* $f(n, X)$ *satisfies*

$$\|f(n, X)\| \geqq m(\|X\|)$$

for all $n \geqq n_0$, *and if the equilibrium* $X = 0$ *of the difference equation* (1) *is uniformly-asymptotically stable, then there exists a real scalar function* $V(n, X)$ *for which the following hold on* D_{n_0R}:

(a) $V(n, X)$ *is positive definite*;

(b) $V(n, X)$ *is decrescent*;

(c) $\Delta V(n, X)$ *is negative definite.*

REFERENCES

[1] W. HAHN, *Stability of Motion*, Academic Press, New York, 1968.

[2] ———, *Über die Anwendung der Methode von Liapunov auf Differenzengleichungen*, Math. Ann., 136 (1958), pp. 430–441.

[3 A. HALANAY, *Quelques questions de la théorie de la stabilité pour les systemes aux differences finies*, Arch. Rational Mech. Anal., 12 (1963), pp. 150–154.

[4] J. HURT, *Some stability theorems for ordinary difference equations*, SIAM J. Numer. Anal., 4 (1967), pp. 582–596.

[5] R. E. KALMAN AND J. E. BERTRAM, *Control analysis and design via the "second method" of Liapunov. II: Discrete-time systems*, ASME J. Basic Engrg. Ser D, 82 (1960), pp. 394–400.

[6] J. L. MASSERA, *Contributions to Stability Theory*, Ann. of Math., 64 (1956), pp. 182–206.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

LINEAR BOUNDED PHASE COORDINATE CONTROL PROBLEMS UNDER CERTAIN REGULARITY AND NORMALITY CONDITIONS*

N. MINAMIDE AND K. NAKAMURA†

Abstract. In this paper, function space bounded phase coordinate control problems are considered by a functional analysis approach. Concepts of regularity and normality are defined and under these conditions, existence and uniqueness of solutions are discussed. Complete characterization of the solution is given in terms of a hyperplane. Furthermore, the relation of the normality condition to a function space version of the bang-bang steering principle is pointed out.

1. Introduction. In recent years, considerable attention has been focused upon the method of functional analysis in the study of optimal control problems which are, in many cases, describable in terms of the optimization of functionals on Banach spaces. This functional analysis method, though applicable to a wide range of problems, seems to be best suited for the investigation of optimization problems arising from linear control systems, since linearity plays an essential role in functional analysis.

In the articles [9] and [10], W. A. Porter formulated Neustadt's minimum effort control problem [8] in Banach space and presented the complete analysis of the abstract problem by using techniques of functional analysis. Also, in [12], a related Banach space minimization problem was considered.

In the present paper, we shall formulate and solve the abstract version of the corresponding bounded phase control problems. Specifically, let X, Y and Z be real Banach spaces. Let $S: X \to Y$ and $T: X \to Z$ be bounded linear transformations.

Problem I. With T onto, $\xi \in Y$ and $\eta \in Z$, find an element, called an optimal solution, (if one exists) $u \in X$ satisfying the constraints $\eta = Tu$ and $\|\xi - Su\| \leqq \varepsilon$ ($\varepsilon > 0$) which minimizes $\|u\|$.

Problem II. With S and T into, find an element (if one exists) $u \in \rho U_X = \{u \mid \|u\| \leqq \rho, u \in X\}$ satisfying $\|\xi - Su\| \leqq \varepsilon$ which minimizes $\|\eta - Tu\|$. If, in Problem I, Z reduces to a trivial Banach space, then there results the following one: $\min \{\|u\| \mid \|\xi - Su\| \leqq \varepsilon, u \in X\}$, which, when the transformation S has dense but not closed range, may provide the natural setting to the minimum effort problem. Problem II has been studied by Gindes [4]. However, it seems that in his argument, a possible mistake was made due to the lack of a necessary hypothesis. In this study, we shall not only touch on this point, but shall also show that the optimal solution is of bang-bang type under certain normality conditions.

2. Some preliminaries. Let us introduce notations and conventions adhered to throughout the paper. Let B be a real Banach space. Let C and D be two sets in B. By the vector sum $C + D$ is meant $C + D = \{c + d | c \in C, d \in D\}$, by $\text{int}(C)$ the interior of C, by ∂C the boundary of C and by $C \times D$ the rectangular set, i.e., $C \times D = \{(c, d) | c \in C, d \in D\}$. Let B' be the conjugate of B. For each $\phi \in B'$, suppose that there exists a vector $x \in U_B$, a closed unit ball in B, such that

* Received by the editors May 7, 1970, and in revised form November 20, 1970.

† Faculty of Engineering, Automatic Control Laboratory, Nagoya University, Furo-Cho, Chikusa-Ku, Nagoya, Japan.

$\langle x, \phi \rangle = \|\phi\|$. Here, $\langle x, \phi \rangle$ denotes the value of a linear functional $\phi \in B'$ at a point $x \in X$. The set of all such vectors x in U_B is called an extremal of ϕ and is denoted by $\bar{\phi}$ (see [9]). For convenience, we sometimes identify a suitable element $x \in \bar{\phi}$ with the set $\bar{\phi}$. It will be obvious from the context whether $\bar{\phi}$ indicates the member or the set. If, for example, $\phi = 0$, then $\bar{\phi}$ denotes a suitable element in U_B, or the set U_B itself. Note that if $\phi \neq 0$, then $\bar{\phi} \subset \partial U_B$.

A convex body is a convex set having a nonempty interior. A convex body K in a Banach space B is called smooth if at each of its boundary points, there is a unique hyperplane of support of K. Also, a convex body K in B is called rotund if K contains no straight-line segments in its boundary (see [14]). A Banach space B is called smooth or rotund according as its unit ball is smooth or rotund. Note that there exists at most one extremal $\bar{\phi}$ of $\phi (\neq 0) \in B'$ if B is rotund.

3. Minimum effort control problem with bounded phase coordinate. In this section, we shall consider Problem I in which T is assumed to be an onto mapping. The methods used in this and the next section are closely related mainly to [11] and others [4], [7].

Let $\hat{S}$ be a linear mapping of X into $Y \times Z$ defined by $\hat{S}: u \to (Su, Tu)$, where $Y \times Z$ denotes a product Banach space equipped with the usual product topology. Let $\hat{S}(U_X)$ denote the image of the unit ball U_X under $\hat{S}$. Motivated by the geometrical interpretation of Problem I, we shall examine the properties of the set $\{\alpha \hat{S}(U_X) + (\varepsilon U_Y \times \{0\})\} = C_\varepsilon(\alpha, 0)$ for $\alpha > 0$ (cf. Porter [9]). Let us begin by introducing the following definition.

DEFINITION. We shall say that a pair (ξ, η) is *regular* if there exists at least one element $u \in X$ satisfying the constraint $\eta = Tu$ and the strict inequality $\|\xi - Su\| < \varepsilon$.

Note that if $\hat{S}$ has dense range, an arbitrary pair (ξ, η) in $Y \times Z$ is regular.

LEMMA 3.1. *The set $C_\varepsilon(\alpha, 0)$ is a convex body.*

Proof. The lemma is an easy consequence of the assumption that T is an onto mapping and the interior mapping principle (see [3, p. 55]).

LEMMA 3.2. *Suppose that $(\xi, \eta) \in \partial C_\varepsilon(\alpha, 0)$ is a regular pair. Then any hyperplane $(\phi_1, \phi_2)(\neq 0) \in (Y \times Z)'$ of support of $C_\varepsilon(\alpha, 0)$ at (ξ, η) satisfies*

$$\langle (\xi, \eta), (\phi_1, \phi_2) \rangle \geqq \alpha \|S'\phi_1 + T'\phi_2\| + \varepsilon \|\phi_1\|, \tag{3.1}$$

$$\|S'\phi_1 + T'\phi_2\| \neq 0, \tag{3.2}$$

where S' denotes the conjugate of S.

Proof. By Lemma 3.1 and the Hahn–Banach theorem [3, p. 58], such a hyperplane as stated in the lemma exists:

$$\langle (\xi, \eta), (\phi_1, \phi_2) \rangle \geqq \langle \alpha(Su, Tu) + \varepsilon(y, 0), (\phi_1, \phi_2) \rangle \qquad \text{for all } u \in U_X, \quad y \in U_Y.$$

Hence taking the supremum of the right side yields (3.1). To see (3.2), suppose to the contrary that $S'\phi_1 + T'\phi_2 = 0$. Then we necessarily have $\phi_1 \neq 0$ and, for all $u \in T^{-1}(\eta) \equiv \{u | \eta = Tu, u \in X\}$,

$$\begin{aligned} \|\xi - Su\| \, \|\phi_1\| &\geqq \langle \xi - Su, \phi_1 \rangle = \langle \xi, \phi_1 \rangle + \langle u, T'\phi_2 \rangle \\ &= \langle (\xi, \eta), (\phi_1, \phi_2) \rangle \geqq \varepsilon \|\phi_1\|. \end{aligned} \tag{3.3}$$

Hence,

$$\|\xi - Su\| \geqq \varepsilon \qquad \text{for all } u \in T^{-1}(\eta), \tag{3.4}$$

which contradicts the regularity of the pair (ξ, η).

The following lemma lists one property of the set $C_\varepsilon(\alpha, 0)$.

LEMMA 3.3. *Let $(\xi, \eta) \in \partial C_\varepsilon(\alpha, 0)$ be a regular pair. Then for all $u \in X$ satisfying $\eta = Tu$ and $\|\xi - Su\| \leqq \varepsilon$, we have*

$$\|u\| \geqq \alpha. \tag{3.5}$$

Proof. Let (ϕ_1, ϕ_2) be the hyperplane in Lemma 3.2. Then for all $u \in T^{-1}(\eta)$, we have

$$\begin{aligned} \|u\| \, \|S'\phi_1 + T'\phi_2\| + \|\xi - Su\| \, \|\phi_1\| &\geqq \langle u, S'\phi_1 + T'\phi_2 \rangle + \langle \xi - Su, \phi_1 \rangle \\ &= \langle (\xi, \eta), (\phi_1, \phi_2) \rangle \\ &\geqq \alpha \|S'\phi_1 + T'\phi_2\| + \varepsilon \|\phi_1\|. \end{aligned} \tag{3.6}$$

Hence,

$$(\|u\| - \alpha)\|S'\phi_1 + T'\phi_2\| \geqq (\varepsilon - \|\xi - Su\|)\|\phi_1\| \qquad \text{for all } u \in T^{-1}(\eta). \tag{3.7}$$

Since we have $\|S'\phi_1 + T'\phi_2\| \neq 0$, this proves the lemma.

LEMMA 3.4. *Suppose that (ξ, η) is regular and that $(\xi, \eta) \in \partial C_\varepsilon(\alpha, 0)$. Then for all $\hat{\alpha} > \alpha$, we have*

$$(\xi, \eta) \in \operatorname{int} \{C_\varepsilon(\hat{\alpha}, 0)\} \subset C_\varepsilon(\hat{\alpha}, 0).$$

Proof. We first note that the assumption $(\xi, \eta) \in \partial C_\varepsilon(\alpha, 0)$ and continuity of the linear form $\langle \, , \rangle$ imply

$$\langle (\xi, \eta), (\psi_1, \psi_2) \rangle \leqq \alpha \|S'\psi_1 + T'\psi_2\| + \varepsilon \|\psi_1\| \qquad \text{for all } (\psi_1, \psi_2) \in (Y \times Z)'. \tag{3.8}$$

Suppose now that the conclusion of the lemma is false and that there exists an $\hat{\alpha} > \alpha$ such that $(\xi, \eta) \notin \operatorname{int} \{C_\varepsilon(\hat{\alpha}, 0)\}$. Then a separating hyperplane $(\phi_1, \phi_2)(\neq 0) \in (Y \times Z)'$ exists:

$$\langle (\xi, \eta), (\phi_1, \phi_2) \rangle \geqq \hat{\alpha} \|S'\phi_1 + T'\phi_2\| + \varepsilon \|\phi_1\|, \tag{3.9}$$

which, combined with the result of Lemma 3.2, contradicts (3.8).

Combining Lemmas 3.3 and 3.4, we have the following theorem.

THEOREM 3.1. *Problem* I *has a solution for each regular pair $(\xi, \eta) \in \partial C_\varepsilon(\alpha, 0)$ if and only if $(\xi, \eta) \in C_\varepsilon(\alpha, 0)$.*

Theorem 3.1 indicates that the existence of solutions to Problem I depends upon whether or not the set $C_\varepsilon(\alpha, 0)$ is closed in $Y \times Z$. We now state sufficient conditions to guarantee this situation.

COROLLARY (cf. [11]). *Suppose that either of the following holds:*

(i) *X is a reflexive Banach space.*

(ii) *There exist normed spaces X_1, Y_1, Z_1 and linear transformations S_1, T_1 such that $X = X_1'$, $Y = Y_1'$, $Z = Z_1'$, $S = S_1'$ and $T = T_1'$, respectively.*

Then Problem I *has a solution for every regular pair.*

Proof. For each regular pair (ξ, η), let α_0 denote the infimum over the set of all real numbers $\alpha \geqq 0$ such that $\{(\xi, \eta) \in C_\varepsilon(\alpha, 0)\}$. It then follows easily that

$(\xi, \eta) \in \partial C_\varepsilon(\alpha_0, 0)$. Hence, it is sufficient to show that $C_\varepsilon(\alpha, 0)$ $(\alpha \geqq 0)$ is closed in $Y \times Z$. We shall do this by assuming (i). The case (ii) may be treated similarly. Note first that $\hat{S}(U_X)$ is weakly compact, being the continuous image of the weakly compact set U_X when Banach spaces X and $Y \times Z$ are equipped with their weak topologies (see [9]). Now, it is known [3, p. 414] that if A and K are closed subsets of an additive topological group G, with K compact, then $A + K$ is closed. Since $U_Y \times \{0\}$ is convex, closed, hence weakly closed in $Y \times Z$ [3, p. 422], it follows that $C_\varepsilon(\alpha, 0)$ is weakly closed, whence closed in $Y \times Z$.

The following lemma characterizes the regular pair (ξ, η) in the dual space.

LEMMA 3.5. *A pair* (ξ, η) *is regular if and only if*

$$\langle (\xi, \eta), (\phi_1, \phi_2) \rangle < \varepsilon \|\phi_1\| \tag{3.10}$$

holds for all $(\phi_1, \phi_2)(\neq 0) \in (Y \times Z)'$ *satisfying* $S'\phi_1 + T'\phi_2 = 0$.

Proof. Only if. If (ξ, η) is a regular pair, then there exists an element $\underline{u} \in X$ such that $\|\xi - S\underline{u}\| < \varepsilon$ and $\eta = T\underline{u}$. Hence it follows from Lemma 3.4 that for $\alpha > \|\underline{u}\|$, $(\xi, \eta) \in \operatorname{int}\{C_\varepsilon(\alpha, 0)\}$. Therefore, for all $(\psi_1, \psi_2)(\neq 0) \in (Y \times Z)'$, we have

$$\langle (\xi, \eta), (\psi_1, \psi_2) \rangle < \alpha \|S'\psi_1 + T'\psi_2\| + \varepsilon \|\psi_1\|, \tag{3.11}$$

from which (3.10) follows.

If. Suppose that (ξ, η) is not a regular pair, i.e., for all u satisfying $\eta = Tu$, we have $\|\xi - Su\| \geqq \varepsilon$. It then follows easily that (ξ, η) cannot be an interior point of $\{\hat{S}(X) + (\varepsilon U_Y \times \{0\})\}$. Hence there exists a separating hyperplane $(\phi_1, \phi_2) \in (Y \times Z)'$ such that

$$\langle (\xi, \eta), (\phi_1, \phi_2) \rangle \geqq \langle u, S'\phi_1 + T'\phi_2 \rangle + \varepsilon \|\phi_1\| \qquad \text{for all } u \in X, \tag{3.12}$$

which, in turn, implies $S'\phi_1 + T'\phi_2 = 0$ and $\langle (\xi, \eta), (\phi_1, \phi_2) \rangle \geqq \varepsilon \|\phi_1\|$. But this contradicts (3.10).

We now state the main result in this section.

THEOREM 3.2. *Suppose that* (ξ, η) *is a regular pair, and that either* (i) *or* (ii) *in the corollary to Theorem* 3.1 *holds. Then an optimal solution* u_0 *of Problem* I *exists, and is necessarily of the form*:

$$u_0 = \frac{\langle (\xi, \eta), (\phi_1, \phi_2) \rangle - \varepsilon \|\phi_1\|}{\|S'\phi_1 + T'\phi_2\|} \overline{(S'\phi_1 + T'\phi_2)}, \tag{3.13}$$

where $(\phi_1, \phi_2) \in (Y \times Z)'$ *of norm* 1 *solves either of the following*:

$$\begin{cases} \xi = \dfrac{\langle (\xi, \eta), \phi_1, \phi_2) \rangle - \varepsilon \|\phi_1\|}{\|S'\phi_1 + T'\phi_2\|} S\overline{(S'\phi_1 + T'\phi_2)} + \varepsilon \bar{\phi}_1, & \text{(3.14a)} \\ \eta = \dfrac{\langle (\xi, \eta), (\phi_1, \phi_2) \rangle - \varepsilon \|\phi_1\|}{\|S'\phi_1 + T'\phi_2\|} T\overline{(S'\phi_1 + T'\phi_2)}, & \text{(3.14b)} \end{cases}$$

$$\max_{\|S'\phi_1 + T'\phi_2\| \neq 0} \left\{ \frac{\langle (\xi, \eta), (\phi_1, \phi_2) \rangle - \varepsilon \|\phi_1\|}{\|S'\phi_1 + T'\phi_2\|} \right\}. \tag{3.15}$$

Conversely, if (ϕ_1, ϕ_2) *of norm* 1 *solves either of the above conditions, then the suitable element* $u_0 \in \{(\langle(\xi, \eta), (\phi_1, \phi_2)\rangle - \varepsilon\|\phi_1\|)/\|S'\phi_1 + T'\phi_2\|\}\overline{(S'\phi_1 - T'\phi_2)}$ *is optimal. Furthermore, if* X *is rotund, the solution is unique.*

Proof. Suppose that $u_0(\neq 0)$ is an optimal solution, and we show (3.13)–(3.15). u_0 thus satisfies $\|\xi - Su_0\| \leqq \varepsilon$ and $\eta = Tu_0$. It further follows that $(\xi, \eta) \in \partial C_\varepsilon(\alpha_0, 0)$, where we put $\|u_0\| = \alpha_0$. Let (ϕ_1, ϕ_2) be a hyperplane of support of $C_\varepsilon(\alpha_0, 0)$ at (ξ, η). We then have, by Lemma 3.2, $S'\phi_1 + T'\phi_2 \neq 0$ and

$$\langle(\xi, \eta), (\phi_1, \phi_2)\rangle \geqq \alpha_0\|S'\phi_1 + T'\phi_2\| + \varepsilon\|\phi_1\|. \tag{3.16}$$

On the other hand, we have

$$\begin{aligned}\langle(\xi, \eta), (\phi_1, \phi_2)\rangle &= \langle \xi - Su_0, \phi_1\rangle + \langle u_0, S'\phi_1 + T'\phi_2\rangle \\ &\leqq \alpha_0\|S'\phi_1 + T'\phi_2\| + \varepsilon\|\phi_1\|.\end{aligned} \tag{3.17}$$

Hence, we conclude that

$$u_0 = \alpha_0\overline{(S'\phi_1 + T'\phi_2)}, \tag{3.18}$$

$$\xi - Su_0 = \varepsilon\bar{\phi}_1, \tag{3.19}$$

$$\alpha_0 = \frac{\langle(\xi, \eta), (\phi_1, \phi_2)\rangle - \varepsilon\|\phi_1\|}{\|S'\phi_1 + T'\phi_2\|}. \tag{3.20}$$

These relations yield (3.13) and (3.14). To see (3.15), note that by (3.8),

$$\langle(\xi, \eta), (\psi_1, \psi_2)\rangle \leqq \alpha_0\|S'\psi_1 + T'\psi_2\| + \varepsilon\|\psi_1\| \qquad \text{for all } (\psi_1, \psi_2) \in (Y \times Z)'.$$

Hence, for all $S'\psi_1 + T'\psi_2 \neq 0$, we have

$$\alpha_0 \geqq \frac{\langle(\xi, \eta), (\psi_1, \psi_2)\rangle - \varepsilon\|\psi_1\|}{\|S'\psi_1 + T'\psi_2\|} \tag{3.21}$$

which, in view of (3.20), yields (3.15).

Conversely, suppose that (ϕ_1, ϕ_2) of norm 1 solves either of conditions (3.14) and (3.15). Let us first consider conditions (3.14). Set

$$\alpha_0 = \frac{\langle(\xi, \eta), (\phi_1, \phi_2)\rangle - \varepsilon\|\phi_1\|}{\|S'\phi_1 + T'\phi_2\|}.$$

It then follows from (3.14) and the equality

$$\begin{aligned}\langle(\xi, \eta), (\phi_1, \phi_2)\rangle &= \alpha_0\|S'\phi_1 + T'\phi_2\| + \varepsilon\|\phi_1\| \\ &= \sup_{\|u\| \leqq 1,\ \|y\| \leqq 1} \{\langle\alpha_0(Su, Tu) + (\varepsilon y, 0), (\phi_1, \phi_2)\rangle\}\end{aligned} \tag{3.22}$$

that $(\xi, \eta) \in C_\varepsilon(\alpha_0, 0) \cap \partial C_\varepsilon(\alpha_0, 0)$. Hence, by Lemma 3.3, $u_0 = \alpha_0\overline{(S'\phi_1 + T'\phi_2)}$ is an optimal solution. Next, consider (3.15). In this case, we have, with α_0 defined as before,

$$\langle(\xi, \eta), (\psi_1, \psi_2)\rangle \leqq \alpha_0\|S'\psi_1 + T'\psi_2\| + \varepsilon\|\psi_1\| \tag{3.23}$$

for all $(\psi_1, \psi_2) \in (Y \times Z)'$ satisfying $S'\psi_1 + T'\psi_2 \neq 0$. But if $S'\psi_1 + T'\psi_2 = 0$, we have, by Lemma 3.5, $\langle(\xi, \eta), (\psi_1, \psi_2)\rangle < \varepsilon\|\psi_1\|$. Hence (3.23) holds for all $(\psi_1, \psi_2) \in (Y \times Z)'$. This, in turn, implies $(\xi, \eta) \in C_\varepsilon(\alpha_0, 0) \cap \partial C_\varepsilon(\alpha_0, 0)$, with (ϕ_1, ϕ_2)

defining a hyperplane of support of $C_\varepsilon(\alpha_0, 0)$ at (ξ, η). Let $u_0 \in \alpha_0 U_X$ and $y_0 \in \varepsilon U_Y$ be any preimage of (ξ, η) so that $(\xi, \eta) = \hat{S}u_0 + (y_0, 0)$. It then follows from Lemma 3.3 and (3.18) that u_0 is an optimal solution and $u_0 \in \alpha_0(\overline{S'\phi_1 + T'\phi_2})$.

Finally, it remains to be shown that u_0 is unique if X is rotund. To this end, let u_0 and u_1 be two solutions. Then $\|u_0\| = \|u_1\| = \alpha_0$ and from (3.16) and (3.17), we have

$$\langle u_0, S'\phi_1 + T'\phi_2\rangle = \langle u_1, S'\phi_1 + T'\phi_2\rangle = \alpha_0\|S'\phi_1 + T'\phi_2\|. \tag{3.24}$$

In other words, the hyperplane $S'\phi_1 + T'\phi_2 \neq 0$ supports $\alpha_0 U_X$ at u_0 and u_1. This implies $u_0 = u_1$ by rotundity of X.

COROLLARY 1. *Suppose that (ξ, η) is a regular pair. Then the following duality relation holds*:

$$\sup_{\|S'\phi_1 + T'\phi_2\| \neq 0} \left\{ \frac{\langle (\xi, \eta), (\phi_1, \phi_2)\rangle - \varepsilon\|\phi_1\|}{\|S'\phi_1 + T'\phi_2\|} \right\} = \inf\{\|u\| \mid \|\xi - Su\| \leqq \varepsilon, \eta = Tu, u \in X\}.$$

COROLLARY 2. *(ϕ_1, ϕ_2) defines a hyperplane of support of $C_\varepsilon(\alpha, 0)$ at (ξ, η) if and only if the vector (ϕ_1, ϕ_2) solves either of the following*:

(i)
$$\begin{cases} \xi = \alpha S(\overline{S'\phi_1 + T'\phi_2}) + \varepsilon\bar{\phi}_1, \\ \eta = \alpha T(\overline{S'\phi_1 + T'\phi_2}), \end{cases}$$

(ii)
$$\max_{\|S'\phi_1 + T'\phi_2\| \neq 0} \left\{ \frac{\langle (\xi, \eta), (\phi_1, \phi_2)\rangle - \varepsilon\|\phi_1\|}{\|S'\phi_1 + T'\phi_2\|} \right\} = \alpha.$$

COROLLARY 3. *Suppose that $(\xi, \eta) \in \partial C_\varepsilon(\alpha, 0)$ is a regular pair, and that Banach spaces X and Y are both smooth. Then there are at most two hyperplanes $(0, \phi_2)$ and $(\phi_1(\neq 0), \phi_2)$ of support of $C_\varepsilon(\alpha, 0)$ at (ξ, η).*

Proof. In proving the theorem, we have shown that ϕ_1 and $S'\phi_1 + T'\phi_2(\neq 0)$ define support hyperplanes of εU_Y and $\alpha_0 U_X$ at $\xi - Su_0$ and u_0, respectively. Hence, by noting that T' is one-to-one, the stated result follows.

COROLLARY 4. *Suppose that (ξ, η) is a regular pair. Then the unique solution to the Hilbert space version of Problem* I *is given by*

$$u_0 = \begin{cases} T^*(TT^*)^{-1}\eta & \text{if } \|ST^*(TT^*)^{-1}\eta - \xi\| \leqq \varepsilon, \\ (\lambda I + S^*S)^{-1}S^*\xi - (\lambda I + S^*S)^{-1}T^*\{T(\lambda I + S^*S)^{-1}T^*\}^{-1} & \\ \quad \cdot \{T(\lambda I + S^*S)^{-1}S^*\xi - \eta\} & \text{if } \|ST^*(TT^*)^{-1}\eta - \xi\| > \varepsilon, \end{cases}$$

where T^* denotes the adjoint of T, and $\lambda > 0$ is a constant uniquely determined by $\|Su_0 - \xi\| = \varepsilon$.

Proof. Note that Hilbert spaces are rotund and smooth, so that there exists a unique extremal $\bar{\phi}$ given by $\bar{\phi} = \phi/\|\phi\|$. This corollary follows from (3.13), (3.14), Corollary 3 and the next lemma (cf. [6]).

LEMMA 3.6. *Let (ξ, η) be a regular pair and suppose that the inequality*

$$\inf_{\substack{\|\xi - Su\| \leqq \varepsilon \\ \eta = Tu}} \|u\| \equiv \alpha_0 > \inf_{\eta = Tu} \|u\| \equiv \alpha_1 \tag{3.25}$$

holds. Then the hyperplane $(\phi_1, \phi_2)(\neq 0) \in (Y \times Z)'$ *of support of* $C_\varepsilon(\alpha_0, 0)$ *at* (ξ, η) *satisfies* $\phi_1 \neq 0$.

Proof. Suppose to the contrary that $\phi_1 = 0$. Then, from (3.1), we have

$$\langle \eta, \phi_2 \rangle \geqq \alpha_0 \| T'\phi_2 \| . \tag{3.26}$$

That is, $\eta \notin \operatorname{int} \{\alpha_0 T(U_X)\}$. But this, in turn, contradicts (3.25) (cf. [9]).

4. Minimization problem with bounded phase coordinate. In the preceding section, the function space version of the minimum effort control problem with bounded phase coordinate was studied. Use of the set $C_\varepsilon(\alpha, 0)$ directly led to the main results: existence theorem, necessary and sufficient conditions, uniqueness theorem for optimality. Attention now turns to the investigation for Problem II. We shall consider, in the present setting, the set $C_\varepsilon(\rho, \alpha) = \{\rho \hat{S}(U_X) + (\varepsilon U_Y \times \alpha U_Z)\}$, $\varepsilon > 0$, $\alpha > 0$. Most of the arguments we develop can parallel those of the preceding section.

DEFINITION 4.1. We shall say that $\xi \in Y$ is (ε, ρ)-*regular* (with respect to S) if there exists at least one element $u \in \rho U_X$ which satisfies $\| \xi - Su \| < \varepsilon$.

LEMMA 4.1. *Let* ξ *be an* (ε, ρ)-*regular element and suppose that* $(\xi, \eta) \in \partial C_\varepsilon(\rho, \alpha)$. *Then any hyperplane* $(\phi_1, \phi_2)(\neq 0)$ *of support of* $C_\varepsilon(\rho, \alpha)$ *at* (ξ, η) *satisfies*

$$\langle (\xi, \eta), (\phi_1, \phi_2) \rangle \geqq \rho \| S'\phi_1 + T'\phi_2 \| + \varepsilon \| \phi_1 \| + \alpha \| \phi_2 \| , \tag{4.1}$$

$$\| \phi_2 \| \neq 0. \tag{4.2}$$

Proof. It is easy to see (4.1). Hence, we shall show (4.2) by contradiction. Suppose that $\phi_2 = 0$. Then we have $\phi_1 \neq 0$ and, for all $u \in X$,

$$\| u \| \, \| S'\phi_1 \| + \| \xi - Su \| \, \| \phi_1 \| \geqq \langle \xi, \phi_1 \rangle \geqq \rho \| S'\phi_1 \| + \varepsilon \| \phi_1 \| .$$

Hence

$$(\| \xi - Su \| - \varepsilon) \| \phi_1 \| \geqq (\rho - \| u \|) \| S'\phi_1 \| \geqq 0 \qquad \text{for all } u \in \rho U_X .$$

This is contradictory to the assumption.

COROLLARY (cf. [2]). ξ *is an* (ε, ρ)-*regular element if and only if*

$$\langle \xi, \phi \rangle < \rho \| S'\phi \| + \varepsilon \| \phi \| \qquad \textit{for all } \phi(\neq 0) \in Y'.$$

Remark. It is easy to see that unless ξ is an (ε, ρ)-regular element, a hyperplane $(\phi_1, \phi_2)(\neq 0) \in (Y \times Z)'$ with $\phi_2 = 0$ does exist which supports $C_\varepsilon(\rho, \alpha)$ at (ξ, η), and yet the problem is well-posed. This point seems to be overlooked in a recent paper by Gindes [4]. The requirement of the regularity condition certainly excludes such a pathological situation as the above lemma indicates.

Throughout the section, we shall assume that ξ is an (ε, ρ)-regular element with respect to S.

LEMMA 4.2. *Suppose that* $(\xi, \eta) \in \partial C_\varepsilon(\rho, \alpha)$. *Then for all* $u \in \rho U_X$ *satisfying* $\| \xi - Su \| \leqq \varepsilon$, *we have*

$$\| \eta - Tu \| \geqq \alpha. \tag{4.3}$$

Proof. With (ϕ_1, ϕ_2) defined in the previous lemma, we have, for all $u \in X$,

$$\|u\| \|S'\phi_1 + T'\phi_2\| + \|\xi - Su\| \|\phi_1\| + \|\eta - Tu\| \|\phi_2\|$$
$$\geqq \langle(\xi, \eta), (\phi_1, \phi_2)\rangle \geqq \rho\|S'\phi_1 + T'\phi_2\| + \varepsilon\|\phi_1\| + \alpha\|\phi_2\|.$$

Hence

$$(\|\eta - Tu\| - \alpha)\|\phi_2\| \geqq (\rho - \|u\|)\|S'\phi_1 + T'\phi_2\| + (\varepsilon - \|\xi - Su\|)\|\phi_1\|, \tag{4.4}$$

which, combined with Lemma 4.1, proves the lemma.

The following results are analogous to those in Lemma 3.4 and Theorem 3.1.

LEMMA 4.3. *Suppose that* $(\xi, \eta) \in \partial C_\varepsilon(\rho, \alpha)$. *Then for all* $\hat{\alpha} > \alpha$, *we have*

$$(\xi, \eta) \in \operatorname{int}\{C_\varepsilon(\rho, \hat{\alpha})\} \subset C_\varepsilon(\rho, \hat{\alpha}).$$

THEOREM 4.1. *Problem* II *has a solution for each* $(\xi, \eta) \in \partial C_\varepsilon(\rho, \alpha)$ *if and only if* $(\xi, \eta) \in C_\varepsilon(\rho, \alpha)$.

In order to completely characterize the optimal solution in terms of the hyperplane, we need the following definitions.

DEFINITION 4.2. We shall say that $\eta \in Z$ is *normal* (with respect to $(\hat{S}, \xi, \varepsilon, \rho)$) if either

$$\inf_{\|u\| \leqq \rho} \{\|\eta - Tu\|\} > \inf_{u \in X} \{\|\eta - Tu\|\} \tag{4.5}$$

or

$$\inf_{\|\xi - Su\| \leqq \varepsilon} \{\|\eta - Tu\|\} > \inf_{u \in X} \{\|\eta - Tu\|\} \tag{4.6}$$

holds.

DEFINITION 4.3. We shall say that $\eta \in Z$ is (ε, ρ)-*normal* (with respect to $(\hat{S}, \xi)$) if

$$\inf_{\substack{\|u\| \leqq \rho \\ \|\xi - Su\| \leqq \varepsilon}} \{\|\eta - Tu\|\} > \inf_{\|\xi - Su\| \leqq \varepsilon} \{\|\eta - Tu\|\} \tag{4.7}$$

holds.

LEMMA 4.4. *Suppose that* (ϕ_1, ϕ_2) *supports* $C_\varepsilon(\rho, \alpha)$ *at* (ξ, η). *Then we have* (cf. LaSalle [5], Schmaedeke and Russell [13])

(i) $\|S'\phi_1 + T'\phi_2\| + \|\phi_1\| \neq 0$, *for each normal element* $\eta \in Z$,

(ii) $\|S'\phi_1 + T'\phi_2\| \neq 0$, *for each* (ε, ρ)-*normal element* $\eta \in Z$.

Proof. The proof is similar to that of Lemma 4.1.

We now summarize.

THEOREM 4.2. *Assume that either* (i) *or* (ii) *stated in the corollary to Theorem* 3.1 *holds. Then there exists a solution to Problem* II *for each* (ε, ρ)-*regular element* ξ. *Suppose, further, that* η *is a normal element. Then* u_0 *is an optimal solution if and only if* u_0 *takes the form*

$$u_0 = \rho(\overline{S'\phi_1 + T'\phi_2}), \tag{4.8}$$

where (ϕ_1, ϕ_2) *of norm* 1 *is determined by either of the following*:

$$\xi = \rho S(\overline{S'\phi_1 + T'\phi_2}) + \varepsilon\bar{\phi}_1, \tag{4.9a}$$

$$\eta = \rho T(\overline{S'\phi_1 + T'\phi_2}) + \{(\langle\xi, \phi_1\rangle + \langle\eta, \phi_2\rangle - \rho\|S'\phi_1 + T'\phi_2\| - \varepsilon\|\phi_1\|)/\|\phi_2\|\}\bar{\phi}_2, \tag{4.9b}$$

$$\max_{\|(\phi_1,\phi_2(\neq 0))\| = 1} \left\{ \frac{\langle(\xi, \eta), (\phi_1, \phi_2)\rangle - \rho\|S'\phi_1 + T'\phi_2\| - \varepsilon\|\phi_1\|}{\|\phi_2\|} \right\}. \tag{4.10}$$

In this case, either $u_0 \in \partial\{\rho U_X\}$ *or* $(\xi - Su_0) \in \partial\{\varepsilon U_Y\}$ *holds. Moreover, if* η *is an* (ε, ρ)*-normal element and if* X *is rotund, then* $u_0 \in \partial\{\rho U_X\}$ *is unique.*

4.1. Application to a minimum effort problem. As an application of the theory developed in this section, we shall consider a minimum effort control problem with amplitude constraints.

Let us suppose that a dynamical system is described by the linear differential equation:

$$dx(t)/dt = Ax(t) + Bu(t),$$

where $x(t)$ is an $n \times 1$ state vector, $u(t)$ is an $r \times 1$ control vector, and A, B are constant matrices of appropriate dimensions. A control vector $u(t)$ which satisfies $|u_j(t)| \leqq \rho$, $j = 1, \cdots, r$, will be called admissible. The problem we shall consider is to find an admissible control vector $u(t)$ which drives the system from a given initial state $x(t_0) = x_0$ to an ε-neighborhood of the target state x^d, i.e., $\max_{1 \leqq j \leqq n} |x_j(t_1) - x_j^d| \equiv \|x(t_1) - x^d\| \leqq \varepsilon$, while minimizing the fuel functional

$$I(u) = \int_{t_0}^{t_1} \sum_{j=1}^{r} |u_j(t)|\, dt,$$

where t_0 and t_1 are fixed initial and final times, respectively.

At the outset, we shall make the following (ε, ρ)-regularity assumption.

(A) There exists at least one admissible control $u(t)$ which enforces the system so that $\|x(t_1) - x^d\| < \varepsilon$.

To formulate the problem in function spaces, let us introduce some standard notations:

$L_p(r, [t_0, t_1])$: The space of (equivalence classes of) r-dimensional vector valued functions, defined and integrable (in the sense of Lebesgue) on the interval $[t_0, t_1]$ equipped with the norm

$$\|f\| = \left\{ \int_{t_0}^{t_1} \sum_{j=1}^{r} |f_j(t)|^p\, dt \right\}^{1/p}, \quad f = (f_1, \cdots, f_r), \quad 1 \leqq p \leqq +\infty,$$

where, for $p = +\infty$, the norm represents the essential supremum of f.

$l_\infty(n)$: The n-dimensional vector space equipped with the norm

$$\|x\| = \max_{1 \leqq j \leqq n} |x_j|.$$

$C(r, [t_0, t_1])$: The space of r-dimensional vector-valued continuous functions defined on $[t_0, t_1]$ equipped with the norm

$$\|f\| = \max_{t_0 \leqq t \leqq t_1} \max_{1 \leqq j \leqq r} |f_j(t)|.$$

$NBV(r,[t_0,t_1])$: The space of r-dimensional vector-valued (normalized) functions of bounded variation on $[t_0,t_1]$ equipped with the norm

$$\|f\| = \sum_{j=1}^{r} v(f_j,[t_0,t_1]), \qquad f(t_0)=0,$$

$v(f_j,[t_0,t_1])$ denoting the total variation of f_j on $[t_0,t_1]$ (see [3, p. 241]).
We then specify the basic function spaces and linear operators as follows:

$$X = L_\infty(r,[t_0,t_1]), \quad Y = l_\infty(n), \quad Z = L_1(r,[t_0,t_1]),$$

$$T(Y \to Z): Tu = -u \text{ (the natural embedding of } X \text{ into } Y),$$
$$S(Z \to Y): Su = \int_{t_0}^{t_1} e^{A(t_1-s)}Bu(s)\,ds.$$

By taking $\xi = x^d - e^{A(t_1-t_0)}x_0$ and $\eta = 0$, the problem at hand is seen to be described in terms of Problem II. Note first that, since $L_1(r,[t_0,t_1])$ can continuously and isometrically be embedded into $NBV(r,[t_0,t_1])$, and since $L_\infty(r,[t_0,t_1])$ and $NBV(r,[t_0,t_1])$ can be identified, respectively, with the duals of $L_1(r,[t_0,t_1])$ and $C(r,[t_0,t_1])$, the assumption (ii) in the corollary to Theorem 3.1 is, in this case, satisfied. Thus, by applying Theorem 4.1, we have

$$u_0 = \rho(\overline{S'\phi_1 - \phi_2}) = \{(\langle \xi, \phi_1\rangle - \rho\|S'\phi_1 - \phi_2\| - \varepsilon\|\phi_1\|)/\|\phi_2\|\}\bar{\phi}_2. \tag{4.11}$$

Here $(S'\phi_1)(t) = B^*\, e^{A^*(t_1-t)}\phi_1$ is an analytic function, and the extremals $(\overline{S'\phi_1 - \phi_2}) \in L_\infty(r,[t_0,t_1])$ and $\bar{\phi}_2 \in L_\infty(r,[t_0,t_1])$ are given, respectively, by (cf. [11])

$$(\overline{S'\phi_1 - \phi_2})_j(t) = \begin{cases} \operatorname{sgn}[(S'\phi_1 - \phi_2)_j(t)], & t \in A_j = \{t \in [t_0,t_1] \mid (S'\phi_1 - \phi_2)_j(t) \neq 0\}, \\ |(\overline{S'\phi_1 - \phi_2})_j(t)| \leqq 1, & t \in A_j^c \text{ (the complement of } A_j), \end{cases}$$

$$(\bar{\phi}_2)_j(t) = \begin{cases} (\bar{\phi}_2)_j(t) \geqq 0, & t \in B_j^+ = \{t \in [t_0,t_1] \mid (\phi_2)_j(t) = \|\phi_2\|\}, \\ (\bar{\phi}_2)_j(t) \leqq 0, & t \in B_j^- = \{t \in [t_0,t_1] \mid (\phi_2)_j(t) = -\|\phi_2\|\}, \\ 0, & t \in (B_j^+ \cup B_j^-)^c, \end{cases}$$

where $j = 1, \cdots, r$ and

$$\int_{t_0}^{t_1} \sum_{j=1}^{r} |(\bar{\phi}_2)_j(t)|\,dt = 1.$$

Hence, by (4.11), the optimal solution can be characterized in a more explicit form:

$$(u_0)_j(t) = \begin{cases} \rho \operatorname{sgn}[(S'\phi_1 - \phi_2)_j(t)], & t \in A_j \subset (B_j^+ \cup B_j^-), \\ 0 \leqq (u_0)_j(t) \leqq \rho, & t \in A_j^c \cap B_j^+, \\ -\rho \leqq (u_0)_j(t) \leqq 0, & t \in A_j^c \cap B_j^-, \\ 0, & t \in (B_j^+ \cup B_j^-)^c. \end{cases} \tag{4.12}$$

We shall show that, if the matrix A is nonsingular, then mes $[A_j^c \cap (B_j^+ \cup B_j^-)]$, the measure of the set $A_j^c \cap (B_j^+ \cup B_j^-)$, is zero, and hence the controls $(u_0)_j(t)$, $j = 1, \cdots, r$, are uniquely determined by (4.12) (cf. [1], [5]).

To see this, suppose to the contrary that mes $[A_j^c \cap (B_j^+ \cup B_j^-)]$ is positive for some j, $1 \leqq j \leqq r$. Then, by appealing to analyticity of the function $(S'\phi_1)_j(t)$, we have

$$|(S'\phi_1)_j(t)| = \|\phi_2\| \qquad \text{for all } t \in [t_0, t_1].$$

On the other hand, it can easily be deduced that, A being nonsingular, the function $(S'\phi_1)_j(t)$ is equal to a constant if and only if $(S'\phi_1)_j(t) = 0$ for all $t \in [t_0, t_1]$. But this contradicts $\|\phi_2\| \neq 0$ by Lemma 4.1.

Remark 1. In order that $S'\phi_1 - \phi_2 \neq 0$ for every hyperplane $(\phi_1, \phi_2)(\neq 0)$ of support of $C_\varepsilon(\rho, \alpha)$ at $(\xi, 0)$, it is necessary and sufficient that

$$\min_{\substack{\|u\| \leqq \rho \\ \|\xi - Su\| \leqq \varepsilon}} \|u\| > \inf_{\|\xi - Su\| \leqq \varepsilon} \|u\|.$$

Hence, it follows that if A is nonsingular, the fuel functional $I(u)$ can be made smaller by enlarging the admissible class of control functions so as to include impulses. In this case the optimal control u_0 may consist of impulse functions.

Remark 2. Instead of assuming that A is nonsingular, mes $[A_j^c \cap (B_j^+ \cup B_j^-)] = 0$ can be shown if there exists a control $u_j \in L_\infty(1, [t_0, t_1])$ such that $\int_{t_0}^{t_1} u_j(s)\, ds \neq 0$ and $\int_{t_0}^{t_1} e^{A(t_1 - s)} b_j u_j(s)\, ds = 0$, where b_j is the jth column vector of the driving matrix $B = (b_1, \cdots, b_r)$.

Acknowledgment. The authors wish to thank Mr. T. Matsuo for his useful and helpful suggestions and discussions.

REFERENCES

[1] M. ATHANS AND P. L. FALB, *Optimal Control*, McGraw-Hill, New York, 1966.
[2] R. CONTI, *Contribution to linear control theory*, J. Differential Equations, 1 (1965), pp. 427–445.
[3] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part I: General Theory*, Interscience, New York, 1967.
[4] V. B. GINDES, *Optimal control in a linear system under conflict*, this Journal, 5 (1967), pp. 163–169.
[5] J. P. LASALLE, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1960, pp. 1–24.
[6] N. MINAMIDE AND K. NAKAMURA, *A restricted pseudoinverse and its application to constrained minima*, SIAM J. Appl. Math., 19 (1970), pp. 167–177.
[7] ———, *A minimum cost control problem in Banach space*, J. Math. Anal. Appl., 35 (1971), pp. 222–234.
[8] L. W. NEUSTADT, *Minimum effort control systems*, this Journal, 1 (1962), pp. 16–31.
[9] W. A. PORTER AND J. P. WILLIAMS, *A note on the minimum effort control problem*, J. Math. Anal. Appl., 13 (1966), pp. 257–264.
[10] W. A. PORTER, *On the optimal control of distributive systems*, this Journal, 3 (1966), pp. 466–472.
[11] ———, *Modern Foundations of Systems Engineering*, Macmillan, New York, 1966.
[12] ———, *A minimization problem and its applications to optimal control and system sensitivity*, this Journal, 6 (1968), pp. 303–311.
[13] W. W. SCHMAEDEKE AND D. L. RUSSELL, *Time optimal control with amplitude and rate limited controls*, this Journal, 2 (1965), pp. 373–395.
[14] F. A. VALENTINE, *Convex Sets*, McGraw-Hill, New York, 1964.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

THE GRADIENT PROJECTION METHOD UNDER MILD DIFFERENTIABILITY CONDITIONS*

G. P. McCORMICK† AND R. A. TAPIA‡

Abstract. Consider the sequence obtained by applying the gradient projection method to the problem of minimizing a continuously differentiable functional over a closed convex subset of a real Hilbert space. In this paper we show that any cluster point of this sequence must be a constrained stationary point.

1. Preliminaries. Let H be a real Hilbert space and R the field of real numbers. Given $f: H \to R$ consider the following problem:

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in S, \tag{1}$$

where S is a closed convex subset of H and f is continuously Fréchet differentiable in H. As usual $(\cdot, \cdot)$ denotes the inner product in H and $\nabla f(x)$ the gradient of f at x. It is clear that the gradient operator $\nabla f: H \to H$ is continuous, since the derivative is continuous.

If C is a closed convex subset of H, then P_C will denote the projection operator for C, i.e., $P_C(x) \in C$ and $\|x - P_C(x)\| \leqq \|x - y\|$ for all $x \in H$ and $y \in C$. It is well known that $P_C: H \to C$ is continuous.

For $x \in S$, the set of all $\eta \in H$ for which there exists a $t > 0$ such that $x + t\eta$ is contained in S is called the tangent cone to S at x and is denoted by $A(x)$. The closure of $A(x)$ in H is denoted by $T(x)$. Clearly $T(x)$ is convex, since $A(x)$ is convex. It is elementary that if S is a polyhedral subset of H, i.e., $S = \{x \in H: (x, \delta_1) \geqq b_1, \cdots, (x, \delta_n) \geqq b_n\}$, then $A(x)$ is closed and $P_{A(x)}(-\nabla f(x))$ is the unique direction of steepest descent in S (see Proposition 1). Clearly, a necessary condition that x solve problem (1) is that $P_{T(x)}(-\nabla f(x)) = 0$. Hence we call any $x \in S$ such that $P_{T(x)}(-\nabla f(x)) = 0$ a constrained stationary point of f. It is a straightforward matter to show that if S is a polyhedral subset of R^n, then x is a constrained stationary point of f if and only if it is a Kuhn–Tucker point.

Given $x_0 \in S$, by the gradient projection method for problem (1) we mean the construction of the sequence

$$x_{n+1} = P_S(x_n - t_n \nabla f(x)), \qquad n = 0, 1, 2, \cdots, \tag{2}$$

where t_n is any solution to

$$\text{minimize} \quad f(P_S(x_n - t\nabla f(x_n))) \quad \text{subject to} \quad t \geqq 0. \tag{3}$$

---

* Received by the editors December 3, 1970, and in final revised form July 9, 1971. A part of this work was sponsored by the United States Army under Contract DA-31-124-ARO-D-462 and was performed while both authors were visiting members of the Mathematics Research Center, University of Wisconsin.

† Research Analysis Corporation, McLean, Virginia 22101.

‡ Department of Mathematical Sciences, Rice University, Houston, Texas 77001.

The technique of obtaining a feasible direction by means of a projection was first used by Rosen [13]. Rosen, for polyhedral S, considers a sequence of the form

$$x_{n+1} = x_n + t_n P_n(-\nabla f(x_n)),$$

where P_n is a projection; however, P_n is not the closest point projection defined above. Balakrishnan [1] uses (2) for the special case when S is a sphere. The use of formula (2) for arbitrary closed and convex S was suggested by A. A. Goldstein [6] and then independently the following year by Levitin and Poljak [10]. However, neither Goldstein nor Levitin and Poljak minimize along a ray. Goldstein requires that f be twice differentiable and that the second derivative be uniformly bounded in a given convex set. The step size, t_n in (2), is then determined according to this bound. Using the slightly milder restriction that the first derivative of f be uniformly Lipschitz continuous in the given convex set Levitin and Poljak do essentially the same thing. These differentiability requirements plus the additional requirement that f be bounded below on S (Goldstein) or that S be bounded (Levitin and Poljak) enable them to prove convergence results. It is the purpose of this paper to analyze the behavior of (2) given only that S is closed and convex and ∇f is continuous on S. Hence we cannot choose t_n in (2) as suggested by Goldstein and Levitin and Poljak.

Certainly there are many objections to the use of the gradient projection method as defined above. Except for special cases, e.g., spheres, linear varieties and orthants, P_S in (2) is very difficult to work with. It is not difficult to construct examples where S is compact, e.g., a sphere in R^n, and (3) does not have a minimum; hence the gradient projection method is not defined. However, the gradient projection method should be easy to implement for the class of problems given in Theorem 2, since in this case P_S can be easily calculated and t_n in (2) need only be a local minimizer of (3).

PROPOSITION 1. *Suppose $T(x)$ is the closure of the tangent cone to S at x, $\eta \in T(x)$, $\omega \in H$ and $\|\eta\| \leqq \|P_{T(x)}(\omega)\|$. Then*

$$(\omega, \eta) \leqq \|P_{T(x)}(\omega)\|^2$$

with equality if and only if $\eta = P_{T(x)}(\omega)$.

Proof. The proof follows from the definitions in a straightforward manner using an elementary inequality given in [7, p. 123].

2. Convergence. The proof of the following proposition was suggested to the authors by R. T. Rockafellar.

PROPOSITION 2. *Let $T(x)$ be the closure of the tangent cone to S at x. Then*

$$\frac{d}{dt_+} P_S(x + ty)\Big|_{t=0} = P_{T(x)}(y) \quad \text{for all } y \in H.$$

Proof. It can be assumed without loss of generality that $x = 0$. Then for each $t > 0$ the point

$$P_S(x + ty) - P_S(x) = P_S(ty)$$

is the point of S nearest ty, and this implies by homogeneity of the norm that $t^{-1}P_S(ty)$ is the point of $t^{-1}S$ nearest to $t^{-1}(ty) = y$. Thus

$$[P_S(x + ty) - P_S(x)]/t = P_{t^{-1}S}(y).$$

As $t \downarrow 0$, the closed convex set $t^{-1}S$ increases, the limit set (i.e., the closure of the union for all $t > 0$) being $T(x)$. It then follows easily that the limit of $P_{t^{-1}S}(y)$ (in the strong topology) is $P_{T(x)}(y)$.

COROLLARY. *If f is differentiable at $x \in S$, then*

$$(4)\quad \frac{d}{dt_+} f(P_S(x - t\nabla f(x)))\Big|_{t=0} = (\nabla f(x), P_{T(x)}(-\nabla f(x))) = -\|P_{T(x)}(-\nabla f(x))\|^2.$$

The second equality follows from Proposition 1. We have from (4) that the path of descent taken by (2) is the steepest descent path possible. Hence the gradient projection method preserves this well-known characterizing property of the gradient method. In particular, for polyhedral S we must have

$$P_S(x - t\nabla f(x)) = x + tP_{A(x)}(-\nabla f(x))$$

for $t \geqq 0$ sufficiently small.

THEOREM 1. *Any cluster point of* (2) *must be a constrained stationary point of f.*

Proof. Let $\{x_n\}$ be the sequence given by (2), x^* a cluster point of $\{x_n\}$ and $\{x_k\}$ a subsequence converging to x^*. If $x_{k+1} = x_k$ for some k, then from (4) we are through. Clearly $g(x, t) = P_S(x - t\nabla f(x)) : S \times R \to S$ is continuous since it is the composition of the continuous functions ∇f, P_S, scalar multiplication and vector addition. Let $t^* = \inf_k t_k$. If $t^* = 0$, then by passing to a subsequence if necessary $t_k \to 0$. From the way t_k is chosen in (2),

$$f(P_S(x_k - (t_k + \tau)\nabla f(x_k))) \geqq f(P_S(x_k - t_k\nabla f(x_k))) \quad \text{for } \tau \geqq 0.$$

Letting $k \to \infty$ and using (4) we have

$$\frac{d}{dt_+} f(P_S(x^* - t\nabla f(x^*)))\Big|_{t=0} \geqq 0;$$

hence from (4), x^* is a constrained stationary point. Now assume $t^* > 0$ and x^* is not a constrained stationary point. We have from (4) that $f(g(x^*, t)) < f(g(x^*, 0)) = f(x^*)$ for small $t > 0$. Choose $0 < \tau < t^*$ such that $f(g(x, \tau)) < f(x^*)$ for all $x \in S$ in some neighborhood of x^*. For k sufficiently large x_k will be in this neighborhood and $f(g(x_k, t_k)) \leqq f(g(x_k, \tau)) < f(x^*)$, which is a contradiction. This proves the theorem.

Remark. The proof of this theorem requires the existence of $\varepsilon > 0$ such that the function in (3) is nondecreasing for $t_n \leqq t \leqq t_n + \varepsilon$ and n sufficiently large. This is clearly weaker than requiring t_n in (3) to be a global minimizer, but stronger than requiring t_n to only be a local minimizer. However, in the unconstrained case, i.e., $S = H$, this proof only requires t_n to be a local minimizer; hence we have Curry's result (see [3]).

3. Orthogonal constraints. By the positive cone of the orthogonal set $\{\delta_\alpha : \alpha \in A\} \subset H$ with respect to $\{b_\alpha : \alpha \in A\} \subset R$ we mean $\{x \in H : (x, \delta_\alpha) \geqq b_\alpha, \alpha \in A\}$.

A well-known example of a positive cone is $\{(x_1, \cdots, x_n) \in R^n : x_1 \geqq 0\}$; more generally we have any polyhedral set with orthogonal constraints, e.g., $\{x \in R^n : Bx \geqq b\}$, where the (not necessarily square) matrix B has orthogonal rows. Clearly a positive cone is closed and convex.

THEOREM 2. *If S in problem* (1) *is a positive cone, then any cluster point of* (2) *must be a constrained stationary point of f, even if one only requires t_n to be a local minimizer of* (3).

Proof. In the proof of this theorem we use the fact that

$$\frac{d}{dt_+} P_S(x_n - (t_n + t)\nabla f(x_n))\Big|_{t=0}$$

exists. This fact does not follow from Proposition 2 since $x_n - t_n \nabla f(x_n)$ is not necessarily contained in S. Hence a proof is required. It is interesting to note that for arbitrary closed and convex S the operator P_S is not necessarily one-sided Gâteaux differentiable at $x \notin S$; see [9].

Suppose $S \neq \varnothing$ is the positive cone of $\{\delta_\alpha : \alpha \in A\}$ with respect to $\{b_\alpha : \alpha \in A\}$. We lose no generality by assuming $\|\delta_\alpha\| = 1$ for $\alpha \in A$. Extend $\{\delta_\alpha : \alpha \in A\}$ to $\{\delta_\alpha : \alpha \in B\}$, a maximum orthonormal subset of H. Let $c \vee d$ denote $\max(c, d)$. A straightforward argument using Parseval's identity shows that

$$P_S(x) = \sum_A [(x, \delta_\alpha) \vee b_\alpha]\delta_\alpha + \sum_{B-A} (x, \delta_\alpha)\delta_\alpha. \tag{5}$$

For $\alpha \in A$ let

$$\Delta(\alpha, x, t, \tau) = \frac{(x - (t + \tau)\nabla f(x), \delta_\alpha) \vee b_\alpha - (x - t\nabla f(x), \delta_\alpha) \vee b_\alpha}{\tau},$$

and for $\alpha \in B - A$ let $\Delta(\alpha, x, t, \tau) = -(\nabla f(x), \delta_\alpha)$. Also, let

$$s_\alpha(x, t) = \lim_{\tau \to 0_+} \Delta(\alpha, x, t, \tau) \quad \text{for all } \alpha \in B.$$

It is clear that $s_\alpha(x, t) = 0$ if $\alpha \in A$ and $(x - t\nabla f(x), \delta_\alpha) < b_\alpha$ or $(x - t\nabla f(x), \delta_\alpha) = b_\alpha$ and $(\nabla f(x), \delta_\alpha) \geqq 0$; otherwise $s_\alpha(x, t) = -(\nabla f(x), \delta_\alpha)$.

It is not difficult to see that $|\Delta(\alpha, x, t, \tau)| \leqq |(\nabla f(x), \delta_\alpha)|$ for all $\alpha \in B$. Therefore,

$$[\Delta(\alpha, x, t, \tau) - s_\alpha(x, t)]^2 \leqq 4(\nabla f(x), \delta_\alpha)^2.$$

By Parseval's identity,

$$\|\nabla f(x)\|^2 = \sum_B (\nabla f(x), \delta_\alpha)^2;$$

hence, for $\varepsilon > 0$ we may choose D, a finite subset of B, such that

$$\sum_{B-D} (\nabla f(x), \delta_\alpha)^2 \leqq \varepsilon.$$

Again by Parseval's identity,

$$\left\| \frac{P_S(x - (t+\tau)\nabla f(x)) - P_S(x - t\nabla f(x))}{\tau} - \sum_B s_\alpha(x,t)\delta_\alpha \right\|^2$$
$$= \sum_B [\Delta(\alpha, x, t, \tau) - s_\alpha(x,t)]^2$$
$$\leqq \sum_D [\Delta(\alpha, x, t, \tau) - s_\alpha(x,t)]^2 + 4\varepsilon.$$

Since D is finite we may choose τ small enough so that $s_\alpha(x,t) = \Delta(\alpha, x, t, \tau)$ for all $\alpha \in D$. It follows that

$$\frac{d}{dt_+} P_S(x - t\nabla f(x)) = \sum_B s_\alpha(x,t)\delta_\alpha;$$

hence

$$\frac{d}{dt_+} f(P_S(x - t\nabla f(x))) = (\nabla f(P_S(x - t\nabla f(x))), \sum_B s_\alpha(x,t)\delta_\alpha). \tag{6}$$

Suppose that t_n is only a local minimizer of (3) and $\{x_k\}$ is a subsequence of (2) converging to x^*. The only part of the proof Theorem 1 that is affected by this change is the case when $t_k \to 0$. From (6) and the way t_k is chosen we have

$$(\nabla f(P_S(x_k - t_k\nabla f(x_k))), \sum_B s_\alpha(x_k, t_k)) > 0.$$

Also, from (5) for all $x \in H$ and $\alpha \in A$,

$$(P_S(x), \delta_\alpha) > b_\alpha \Leftrightarrow (x, \delta_\alpha) > b_\alpha. \tag{7}$$

Since $x^* \in S$, $(x^*, \delta_\alpha) \geqq b_\alpha$ for all $\alpha \in A$. For a particular $\alpha \in A$, if $(P_S(x^*), \delta_\alpha) = b_\alpha$ and $(P_S(x^* - \tau\nabla f(x^*)), \delta_\alpha) > b_\alpha$ for some $\tau > 0$, then $(\nabla f(x^*), \delta_\alpha) < 0$. By continuity and (7), for k sufficiently large, $(P_S(x_k - t_k\nabla f(x_k)), \delta_\alpha) > 0$. We may now conclude that $s_\alpha(x^*, 0) \neq 0$ implies $s_\alpha(x_k, t_k) \neq 0$. It follows that if $B_* = \{\alpha \in B : s_\alpha(x^*, 0) \neq 0\}$, then

$$\frac{d}{dt_+} f(P_S(x^* - t\nabla f(x^*)))\Big|_{t=0} = -\lim_{k\to\infty} \sum_{B^*} (\nabla f(P_S(x_k - t_k\nabla f(x_k))), \nabla f(x_k)) \geqq 0.$$

Hence x^* must be a constrained stationary point of f. This proves the theorem.

Remark. Formula (5) gives a constructive method for evaluating $P_S(x)$ which, at least for polyhedral sets, should be easy to implement.

Remark. Property (7) allowed us to relax the method of selecting t_n in (2). It is not difficult to show that this property characterizes orthogonal constraints; hence it applies to polyhedral sets only in this special case.

Acknowledgments. The authors would like to thank Professors A. A. Goldstein, H. Halkin and the referee for many helpful suggestions. They are also indebted to Professor R. T. Rockafellar for the proof of Proposition 2, and to Professor E. H. Zarantonello for allowing them the use of [19].

REFERENCES

[1] A. V. BALAKRISHNAN, *An operator theoretic formulation of a class of control problems and a steepest descent method of solution*, this Journal, 1 (1962), pp. 109–127.

[2] T. A. BROWN, M. JUNCOSA AND V. KLEE, *Invertibly positive linear operators on spaces of continuous functions*, Math. Ann., 183 (1969), pp. 105–114.

[3] H. B. CURRY, *The method of steepest descent for nonlinear minimization problems*, Quart. Appl. Math., 2 (1944), pp. 258–261.

[4] J. W. DANIEL, *Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[5] C. GOFFMAN AND G. PEDRICK, *First Course in Functional Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1965.

[6] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[7] ———, *Constructive Real Analysis*, Harper and Row, New York, 1967.

[8] M. GOLOMB, *Zur Theorie der nichtlinear Integral-gleichungen*, Math. Z., 39 (1934), pp. 45–75.

[9] J. B. KRUSKAL, *Two convex counterexamples: a discontinuous envelope function and a nondifferentiable nearest point mapping*, Proc. Amer. Math. Soc., 23 (1969), pp. 697–203.

[10] E. S. LEVITIN AND B. T. POLJAK, *Constrained minimization methods*, Zh. Vychisl. Mat. i Mat. Fiz., 6 (1965), pp. 787–823 = U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 1–50.

[11] G. P. MCCORMICK, *Anti-zig-zagging by bending*, Management Sci., 15 (1969), pp. 315–320.

[12] G. P. MCCORMICK AND R. A. TAPIA, *The gradient projection method and Curry's theorem*, Rep. 1074, Mathematics Research Center, University of Wisconsin, Madison, 1970.

[13] J. B. ROSEN, *The gradient projection method for nonlinear programming. Part I*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 181–217.

[14] R. A. TAPIA, *Integration and differentiation of nonlinear operators*, Nonlinear Functional Analysis and Applications, L. B. Rall, ed., Academic Press, New York, 1971.

[15] B. Z. VULIKH, *Introduction to the Theory of Partially Ordered Spaces*, Wolters-Noordhoff, Groningen, the Netherlands, 1967.

[16] A. WILANSKY, *Functional Analysis*, Blaisdell, New York, 1964.

[17] P. WOLFE, *On the convergence of gradient methods under constraint*, Res. Paper RC-1752, IBM Watson Research Center, Yorktown Heights, N.Y., 1967.

[18] ———, *Convergence conditions for ascent methods*, SIAM Rev., 11 (1969), pp. 226–234.

[19] E. H. ZARANTONELLO, *Projections on convex sets in Hilbert spaces*, unpublished manuscript.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

PHASE FUNCTION NORM ESTIMATES FOR STABILITY OF SYSTEMS WITH MONOTONE NONLINEARITIES*

MARVIN I. FREEDMAN†

Abstract. A system involving a linear element $G(s)$ and odd monotone nonlinearity $N(\cdot)$ is considered. It is assumed that this system is stable for all $N(r) = lr$, $0 \leqq l \leqq k$. General $N(\cdot)$ satisfying $0 \leqq (N(r) - N(s))/(r - s) \leqq k - \varepsilon$ (for distinct r, s and sufficiently small $\varepsilon > 0$) are considered.

Letting $\Phi(\omega)$ denote the phase function $\arg\{G(i\omega) + 1/k\}$ and $\Phi'(\omega)$ its derivative, stability theorems are proven which involve certain norms of $\Phi(\omega)$ and its derivative $\Phi'(\omega)$. Finally, a recipe is given for generating new stability results with predetermined characteristics.

1. Introduction. Over the past few years there has been a certain interest in developing stability criteria for a system involving a time-invariant open loop closed by a monotone nonlinearity in feedback. (See Fig. 1.) Results on this problem have been obtained by O'Shea [1] and Falb and Zames [2].

Each of these papers proves results of a "multiplier" type. For instance, in [2] Falb and Zames show that for a system consisting of a convolution operator G and an odd monotone nonlinearity $N(\cdot)$ satisfying $0 \leqq (N(r) - N(s))/(r - s) \leqq k - \varepsilon$, stability can be proved if there exists an operator $Z: L_2[0, \infty) \to L_2[0, \infty)$ defined by

$$(Zx)(t) = x(t) + \int_0^t z_1(t - \tau)x(\tau)\, d\tau,$$

where

$$z_1(\cdot) \in L_1[0, \infty), \qquad \int_0^\infty |z_1(t)|\, dt < 1,$$

and satisfying

$$\operatorname{Re} Z(i\omega)[G(i\omega) + 1/k] \geqq \delta > 0$$

for all real ω.

This result therefore depends on the existence of a "multiplier" operator which when combined with $[G(s) + 1/k]$ yields a positive operator.

The usefulness of the multiplier approach is limited by the absence of any explicit method for finding suitable multipliers.

In [3], this author and G. Zames studied the stability of a feedback system involving a linear time-invariant element $G(s)$ coupled with a time-varying gain $n(\cdot)$. The method of study involved the development of a constructive process which allowed for removal of the multiplier function from the final results; these results being geometric in nature.

In [4], this author considered a second stability situation (previously discussed in Brockett and Forys [5]) involving a time-varying gain in feedback. Once again it was possible to eliminate the multiplier dependence from the final answer, this time by bringing into play the phase function $\Phi(\omega) \triangleq \arg\{G(i\omega) + 1/k\}$ and its

* Received by the editors March 19, 1971, and in revised form July 13, 1971.

† Department of Mathematics, Boston University, Boston, Massachusetts 02215.

derivative $\Phi'(\omega)$. (See Fig. 2.) Very roughly, the stability result of [4] said that the rate of change allowable for the gain, with respect to time, should be inversely proportional to the magnitude of $\Phi'(\omega)$, i.e., the angular rate with which the Nyquist plot is swept out with respect to frequency.

In this paper, attention is again focused on the monotone nonlinearity situation, but as in [3] and [4] a method of constructing a suitable multiplier is given and the main results (see Theorems 6.1, 7.3 and 7.5) involve only certain (integral) bounds on $\Phi(\omega) = \arg\{G(i\omega) + 1/k\}$ and $\Phi'(\omega)$. Roughly speaking, the factors entering the stability considerations are (i) the amount by which $|\Phi(\omega)|$ exceeds $\pi/2$, (ii) for what ω-duration does $|\Phi(\omega)|$ exceed $\pi/2$ and (3) the magnitude of $\Phi'(\omega)$ for those frequencies that $|\Phi(\omega)|$ exceeds $\pi/2$ (i.e., how swiftly with respect to frequency does the angle function $\Phi(\omega)$ sweep).

2. Mathematical preliminaries.

DEFINITION 2.1. Let $L_p[0, \infty)$, where $1 \leqq p < \infty$, denote the linear space of real-valued measurable functions $x(\cdot)$ on $[0, \infty)$ with the property that

$$\int_0^\infty |x(t)|^p \, dt < \infty.$$

Let $L_p[0, \infty)$ be normed with the norm

$$\|x(\cdot)\|_p = \left\{ \int_0^\infty |x(t)|^p \, dt \right\}^{1/p}.$$

The spaces $L_p(-\infty, \infty)$ on the interval $(-\infty, \infty)$ are similarly defined.

The definition of the extended space L_{2e} is introduced next (see [6]).

DEFINITION 2.2. Let L_{2e} be the space of real-valued measurable functions $x(\cdot)$ on $[0, \infty)$ satisfying

$$\int_0^T |x(t)|^2 \, dt < \infty \qquad \text{for all } T > 0.$$

2.1. Feedback equations and stability. The feedback system of Fig. 1 will be represented for all $t \geqq 0$ by the nonlinear integral equation

$$e(t) = x(t) - N\left(\int_0^t e(\tau) g(t - \tau) \, d\tau \right) \tag{2.1}$$

or, alternatively, the pair

$$\begin{aligned} e(t) &= x(t) - N(y(t)), \\ y(t) &= \int_0^t e(\tau) g(t - \tau) \, d\tau \end{aligned} \tag{2.2}$$

in which the following assumptions are made.

Assumption 1. $x(\cdot)$ is in $L_2[0, \infty)$. (The function $x(\cdot)$ represents the combined effects of an input and of possible nonzero initial conditions.)

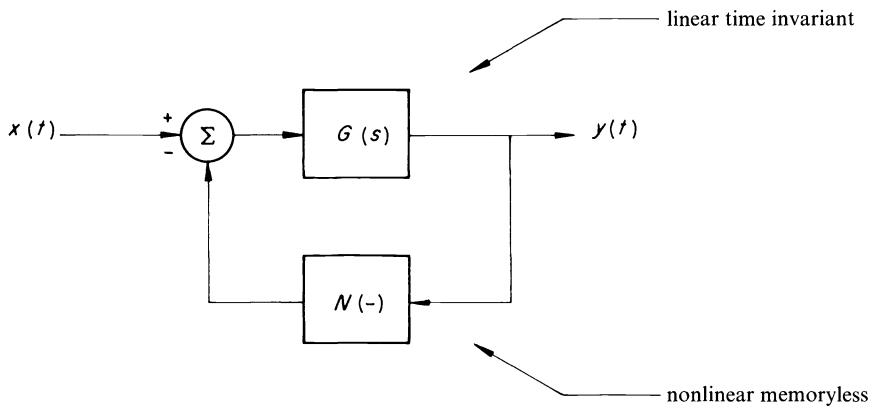


FIG. 1. *A feedback system*

Assumption 2. $g(\cdot)$ is in $L_1[0, \infty)$.

Assumption 3. $N(\cdot)$ is a real-valued function on $(-\infty, \infty)$ with the following properties:

(i) $N(\cdot)$ is odd, i.e., $N(-x) = -N(x)$ (in particular $N(0) = 0$),

(ii) N is monotone nondecreasing, i.e.,

$$(r - s)[N(r) - N(s)] \geqq 0 \qquad \text{for all } r, s \text{ real.}$$

Assumption 4. $e(\cdot)$ (and also $y(\cdot)$) is in L_{2e} (i.e., existence of solutions in L_{2e} for $L_2[0, \infty)$-inputs is being assumed).

DEFINITION 2.3. Feedback system (2.1) will be termed *L_2-stable* if for any pair $(x(\cdot), e(\cdot))$ for which (2.1) and the related Assumptions 1–4 hold, $e(\cdot)$ is in $L_2[0, \infty)$ with $\|e(\cdot)\|_2 \leqq \text{const } \|x(\cdot)\|_2$.

This notion of stability in the context of differential equations implies asymptotic stability ($\lim_{t\to\infty} y(t) = 0$) and, with additional minor assumptions, can be used to show bounded-input bounded-output stability.

2.2. Some special notation.

DEFINITION 2.4. For any $k > 0$ let

$$W_k = \{g(\cdot) \in L_1[0, \infty) | G(i\omega) + 1/l \neq 0 \text{ for } -\infty < \omega < \infty \text{ and all } l, 0 < l \leqq k\},$$

where $G(s)$, the Laplace transform of $g(\cdot)$, is the complex-valued function defined on $\{s | \text{Re}\{s\} \geqq 0\}$ by the integral

$$G(s) = \int_0^\infty e^{-st} g(t)\, dt.$$

Remark 2.5. $g(\cdot) \in W_k$ if and only if the set $\{G(i\omega) | \omega \in (-\infty, \infty)\}$ does not cut the negative real axis from $-\infty$ up to and including the point $-1/k$. The classical Nyquist criterion (see [7]) assures that $g(\cdot)$ in W_k is a necessary and sufficient condition for feedback systems (2.1) to be L_2-stable for all constant gains between 0 and k, i.e., for $N(r) \equiv lr$, where l is a constant $0 \leqq l \leqq k$.

Remark 2.6, *and some special notation.* Given $g(\cdot) \in W_k$, define the phase function

$$\Phi(\omega) \triangleq \arg \{G(i\omega) + 1/k\}.$$

Then since $G(i\omega) + 1/k$ does not cut the negative real axis as ω passes from $-\infty$ to ∞, it follows that $\Phi(\omega)$ is uniquely defined for all ω and takes values in $(-\pi, \pi)$. Since $g(\cdot) \in L_1[0, \infty)$, the Riemann–Lebesgue lemma assures that $\lim_{|\omega| \to \infty} G(i\omega) = 0$ and so $\lim_{|\omega| \to \infty} \arg \{G(i\omega) + 1/k\} = 0$ also.

3. The multiplier algebra.

DEFINITION 3.1. Let $\mathscr{L}$ denote the space of linear maps $\underline{H}: L_{2e}[0, \infty) \to L_{2e}[0, \infty)$ satisfying

$$(\underline{H}x)(t) = h_0 x(t) + \int_0^t x(\tau) h_1(t - \tau)\, d\tau$$

for all $x(\cdot)$ in L_{2e} and all $t \geqq 0$. Here h_0 is a real number and $h_1(\cdot)$ an $L_1[0, \infty)$ function. We write $\{h_0, h_1(\cdot)\}$ for $\underline{H}$ as a mnemonic.

LEMMA 3.2. *$\mathscr{L}$ is a commutative Banach algebra with the norm $\|\cdot\|$ given by*

$$\|\underline{H}\| = |h_0| + \|h_1(\cdot)\|_1 \qquad \text{for } \underline{H} = \{h_0, h_1(\cdot)\}.$$

Proof. If $\underline{G} = \{g_0, g_1(\cdot)\}$ and $\underline{H} = \{h_0, h_1(\cdot)\}$ are two elements of $\mathscr{L}$, then the "product"

$$\underline{G}\underline{H} \triangleq \{g_0 h_0, g_0 h_1(\cdot) + h_0 g_1(\cdot) + g_1(\cdot) * h_1(\cdot)\}$$

lies in $\mathscr{L}$. The $*$ denotes the convolution product; that is,

$$(x_1 * x_2)(t) \triangleq \int_0^\infty x_1(t - \tau) x_2(\tau)\, d\tau.$$

In fact,

$$\begin{aligned} \|\underline{G}\underline{H}\| &= |h_0 g_0| + \|g_0 h_1(\cdot) + h_0 g_1(\cdot) + g_1(\cdot) * h_1(\cdot)\|_1 \\ &\leqq (|g_0| + \|g_1(\cdot)\|_1)(|h_0| + \|h_1(\cdot)\|_1) = \|\underline{G}\|\, \|\underline{H}\|. \end{aligned}$$

Remark 3.3. The algebra $\mathscr{L}$, defined above, is sometimes called "the algebra of causal[1] multipliers."

For $\underline{Z} \in \mathscr{L}$ there is a natural notion of Laplace transform.

DEFINITION 3.4. Given $\underline{Z} \in \mathscr{L}$, $Z = \{z_0, z_1(\cdot)\}$, we define $Z(s)$, called the Laplace transform of $\underline{Z}$, by

$$Z(s) = z_0 + \int_0^\infty e^{-st} z_1(t)\, dt$$

for any s with $\operatorname{Re}\{s\} \geqq 0$.

We shall have use for the following known multiplier stability result. See [1] and [2].

[1] See [6] for definition of causal.

THEOREM 3.5. *Consider the feedback system described by* (2.1) *and pictured in Fig.* 1. *Let Assumptions* 1–4 *hold. Let* $k > 0$ *be a positive number and assume* $g(\cdot) \in W_k$. *Further let* $\varepsilon > 0, 0 \leqq \varepsilon < k$ *and assume* $0 \leqq (N(r) - N(s))/(r - s) \leqq k - \varepsilon$ *for all distinct real numbers* r *and* s. *Then a sufficient condition for this feedback equation to be* L_2*-stable is that there exist a multiplier* $\mathcal{Z}$ *in* $\mathcal{L}$ *of the form* $\{1, z_1(\cdot)\}$ *which satisfies the following two conditions*:

(i) $\|z_1(\cdot)\|_1 < 1$,

(ii) *there exists* $\delta > 0$ *such that*

$$\operatorname{Re}\{Z(i\omega)[G(i\omega) + 1/k]\} \geqq \delta \qquad \text{for all real } \omega.$$

Remark 3.6. A proof of this theorem appears in Falb and Zames [2]. In fact, [2] contains a stronger result than the one we quote here, as *causality* was not required for the multiplier $\mathcal{Z}$. Also, in [2] a result is given for nonlinearities that are not odd. We however, do not make use of these stronger results in this paper.

4. Multipliers with prescribed phase. In a paper of Freedman and Zames [3], the following result was presented.

THEOREM 4.1. *If*

(i) $s(\omega)$ *is a real-valued a.e. differentiable odd function of* ω, *for* $\omega \in (-\infty, \infty)$,

(ii) $s(\omega)$ *and* $s'(\omega)$ *both lie in* $L_2(-\infty, \infty)$,

then

(a) *there is a function* $\lambda(\cdot)$ *in* $L_1(-\infty, \infty)$ *with* $\lambda(t) = 0$ *for* $t < 0$ *and with Laplace transform* $\Lambda(s)$ *satisfying* $\operatorname{Im}\{\Lambda(i\omega)\} = s(\omega)$;

(b) *there is a* $y(\cdot)$ *in* $L_1(-\infty, \infty)$ *with* $y(t) = 0$ *for* $t < 0$, *and with Laplace transform* $Y(s)$ *satisfying* $1 + Y(s) = \exp[\Lambda(s)]$ *for* $\operatorname{Re}\{s\} \geqq 0$;

(c) *if* $-\pi < s(\omega) < \pi$, *there is a* $y(\cdot) \in L_1(-\infty, \infty)$ *with* $y(t) = 0$ *for* $t < 0$, $1 + Y(s) \neq 0$ *in* $\operatorname{Re}\{s\} \geqq 0$ *and* $\arg\{1 + Y(i\omega)\} = s(\omega)$ *for all real* ω.

Remark 4.2. The proof appears in [3] (see also [4]). For our purposes in succeeding sections of this paper, it suffices to recall that in the proof $y(\cdot)$ is the L_1-norm-limit of a sequence $\{y_n(\cdot)\}$, where

$$y_n(t) = \lambda(t) + \frac{(\lambda * \lambda)(t)}{2!} + \cdots + \frac{\overbrace{(\lambda * \lambda * \cdots * \lambda)}^{n}(t)}{n!};$$

the $\lambda(\cdot)$ is the $L_1(-\infty, \infty)$ function with support in $[0, \infty)$ given by

$$\lambda(t) = \begin{cases} 2\phi_0(t), & t \geqq 0, \\ 0, & t < 0, \end{cases}$$

where $\phi_0(t)$ is the inverse limit-in-the-mean Fourier transform of $is(\omega)$.

Remark 4.3. Paraphrasing part (c) of Theorem 4.1, we see that if $s(\omega)$ is any function satisfying conditions (i) and (ii) of Theorem 4.1 with $-\pi < s(\omega) < \pi$ for all real ω, then there exists a unique (causal) multiplier $\mathcal{Z} \in \mathcal{L}$ of the form

$$\mathcal{Z} = \{1, y(\cdot)\} \quad \text{satisfying} \quad \arg\{Z(i\omega)\} = s(\omega)$$

for all ω real.

5. The multiplier norm estimate. As a consequence of the material presented in § 3, it follows that in order to obtain stability results for feedback equation (2.1), it will suffice to find a multiplier $\mathcal{Z} \in \mathscr{L}$ of the form $\mathcal{Z} = \{1, y(\cdot)\}$ with $\|y(\cdot)\|_1 < 1$ which satisfies an estimate of the form $\operatorname{Re}\{Z(i\omega)G(i\omega)\} \geqq 0$ for all real ω.

In this section, by delving more deeply into the material of § 4, we develop a condition on the argument function $s(\omega)$ sufficient to ensure that the element $\mathcal{Z} \in \mathscr{L}$, with $\mathcal{Z} = \{1, y(\cdot)\}$ which satisfies $\arg\{Z(i\omega)\} = s(\omega)$ for all real ω, will additionally satisfy $\|y(\cdot)\|_1 < 1$.

THEOREM 5.1 (norm estimate theorem). *Let $s(\omega)$ be a real-valued a.e. differentiable odd function of ω, for $\omega \in (-\infty, \infty)$. Assume that both $s(\omega)$ and $s'(\omega)$ are in $L_2(-\infty, \infty)$ and also that $-\pi < s(\omega) < \pi$ for all $\omega \in (-\infty, \infty)$. Then a sufficient condition for the multiplier $\mathcal{Z} \in \mathscr{L}$, $\mathcal{Z} = \{1, y(\cdot)\}$ with $\arg\{Z(i\omega)\} = s(\omega)$, to satisfy $\|y(\cdot)\|_1 < 1$ is that*

$$\left(\int_{-\infty}^{\infty} s(\omega)^2\, d\omega\right)^{1/2}\left(\int_{-\infty}^{\infty} s'(\omega)^2\, d\omega\right)^{1/2} < (\log 2)^2. \tag{5.1}$$

In the proof of Theorem 5.1 we utilize the following elementary lemma.

LEMMA 5.2. *For any two positive numbers a and b,*

$$2ab = \min_{\alpha > 0} \{a^2/\alpha + \alpha b^2\}.$$

Proof of Lemma 5.2. For any real α,

$$(\alpha b - a)^2 \geqq 0 \quad \text{so} \quad \alpha^2 b^2 - 2\alpha b a + a^2 \geqq 0.$$

Then, rearranging and dividing by α gives

$$2ab \leqq (1/\alpha)a^2 + \alpha b^2 \qquad \text{provided } \alpha > 0.$$

But clearly if $\alpha = a/b$, we have

$$2ab = (1/\alpha)a^2 + \alpha b^2.$$

Proof of Theorem 5.1. First note that Lemma 5.2 assures that if inequality (5.1) holds, then there exists some $\alpha > 0$ with

$$\frac{1}{\alpha}\int_{-\infty}^{\infty} s(\omega)^2\, d\omega + \alpha \int_{-\infty}^{\infty} s'(\omega)^2\, d\omega < 2(\log 2)^2. \tag{5.2}$$

Actually, if either $\int_{-\infty}^{\infty} s(\omega)^2\, d\omega$ or $\int_{-\infty}^{\infty} s'(\omega)^2\, d\omega$ were zero, the lemma would not apply. Clearly however, if $\int_{-\infty}^{\infty} s(\omega)^2\, d\omega = 0$, then $s(\omega) = 0$ a.e.; and so $\int_{-\infty}^{\infty} s'(\omega)^2\, d\omega = 0$ also, making the estimate (5.2) trivially true. In the case $\int_{-\infty}^{\infty} s'(\omega)^2\, d\omega = 0$, just take for α any sufficiently large number and (5.2) will be assured.

Now, as in Remark 4.2, let $\phi_0(t)$ be the inverse limit-in-the-mean Fourier transform of $is(\omega)$. Then note that

$$\begin{aligned}\int_{-\infty}^{\infty} |\phi_0(t)|\,dt &= \int_{-\infty}^{\infty} \frac{1}{(1+\alpha^2t^2)^{1/2}}(1+\alpha^2t^2)^{1/2}|\phi_0(t)|\,dt \\ &\leqq \left(\int_{-\infty}^{\infty} \frac{dt}{1+\alpha^2t^2}\right)^{1/2}\left(\int_{-\infty}^{\infty}(1+\alpha^2t^2)\,\phi_0(t)^2\,dt\right)^{1/2} \\ &\leqq \left(\frac{\pi}{\alpha}\right)^{1/2}\left(\int_{-\infty}^{\infty}\phi_0(t)^2\,dt + \alpha^2\int_{-\infty}^{\infty} t^2\phi_0(t)^2\,dt\right)^{1/2},\end{aligned}$$

as is seen by an application of the Schwarz inequality.

But by Parseval's theorem,

$$\int_{-\infty}^{\infty}\phi_0(t)^2\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} s(\omega)^2\,d\omega$$

and also

$$\int_{-\infty}^{\infty} t^2\phi_0(t)^2\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} s'(\omega)^2\,d\omega,$$

from Parseval's theorem combined with [9, Thm. 68, p. 92]. Therefore,

$$\int_{-\infty}^{\infty}|\phi_0(t)|\,dt \leqq \left(\frac{\pi}{\alpha}\right)^{1/2}\left(\frac{1}{2\pi}\right)^{1/2}\left(\int_{-\infty}^{\infty}(s(\omega)^2+\alpha^2s'(\omega)^2)\,d\omega\right)^{1/2}. \tag{5.3}$$

Now since

$$\lambda(t) = \begin{cases} 2\phi_0(t), & t \geqq 0, \\ 0, & t < 0, \end{cases}$$

it follows that $\|\lambda(\cdot)\|_1 = \int_{-\infty}^{\infty}|\phi_0(t)|\,dt$, since $\phi_0(t)$ must be an odd function of t.

Rewriting (5.3) now yields

$$\|\lambda(\cdot)\|_1 \leqq \frac{1}{\sqrt{2}}\left(\frac{1}{\alpha}\int_{-\infty}^{\infty} s(\omega)^2\,d\omega + \alpha\int_{-\infty}^{\infty} s'(\omega)^2\,d\omega\right)^{1/2}, \tag{5.4}$$

and combining (5.4) with (5.2) gives the simple estimate

$$\|\lambda(\cdot)\|_1 < \log 2. \tag{5.5}$$

Now $y_n(t) = \lambda(t) + (\lambda * \lambda)(t)/2! + \cdots + \overbrace{(\lambda * \cdots * \lambda)}^{n}(t)/n!$, and so

$$\begin{aligned}\|y_n(\cdot)\|_1 &\leqq \|\lambda(\cdot)\|_1 + \frac{\|\lambda(\cdot)\|_1^2}{2!} + \cdots + \frac{\|\lambda(\cdot)\|_1^n}{n!} \\ &\leqq e^{\|\lambda(\cdot)\|_1} - 1 \quad \text{for each integer } n \geqq 1.\end{aligned}$$

Since $y(\cdot)$ is the L_1-norm-limit of the $y_n(\cdot)$, it follows also that

$$\|y(\cdot)\|_1 \leqq e^{\|\lambda(\cdot)\|_1} - 1.$$

But using (5.5), we see that

$$e^{\|\lambda(\cdot)\|_1} - 1 < e^{\log 2} - 1 < 1.$$

Thus the multiplier $\underset{\sim}{Z} \in \mathscr{L}$ of the form $Z = \{1, y(\cdot)\}$ with $\arg Z(i\omega) = s(\omega)$ will satisfy $\|y(\cdot)\|_1 < 1$ provided estimate (5.1) holds for $s(\omega)$.

Remark 5.3. In essence the result is based on finding a suitable condition on $\lambda(\cdot)$ such that

$$y(\cdot) = \lambda(\cdot) + \frac{(\lambda * \lambda)(\cdot)}{2!} + \cdots$$

will satisfy $\|y(\cdot)\|_1 < 1$. For real numbers the algebraic equation $y = \lambda + \lambda^2/2! + \cdots = e^{\lambda} - 1$ will be satisfied with $|y| < 1$ provided $|\lambda| < \log 2$, surely, but also any negative λ will also assure $|y| < 1$.

In the Banach algebra setting considered here, there seems to be no clear analogue of this latter possibility (i.e., a notion of negativeness on $\lambda(\cdot)$ under which $\|y(\cdot)\|_1 < 1$ will hold) and it is highly questionable whether any such analogue does exist. However, an affirmative answer would broaden the scope of this technique considerably.

6. The main stability theorem. We now are ready to combine the material of §§ 3, 4 and 5 in order to state our main stability result. As in § 2, $\Phi(\omega)$ will denote $\arg\{G(i\omega) + 1/k\}$ (see Fig. 2).

THEOREM 6.1. *Suppose feedback equation* (2.1) (*or equivalently* (2.2)) *and the related Assumptions* 1–4 *hold for a pair* $(x(\cdot), e(\cdot))$. *Let* $k > 0, \varepsilon > 0$ *be given and assume that*:

(i) $0 \leqq (N(r) - N(s))/(r - s) \leqq k - \varepsilon$ *for all real* r, s *with* $r \neq s$,

(ii) $g(\cdot) \in W_k$ (*i.e., the system* (2.1) *is* L_2*-stable when* $N(r) \equiv lr$ *for any* $r, 0 \leqq l \leqq k$.

Further, assume that there exists a real-valued a.e. differentiable function $s(\omega)$ *with* $s(0) = 0$ *defined on* $[0, \infty)$ *and satisfying*:

(a) $-\pi < s(\omega) < \pi$ *for all* $\omega \geqq 0$,

(b) $|s(\omega) + \Phi(\omega)| < \pi/2$ *for all* $\omega \geqq 0$,

(c) $s(\omega)$ *and* $s'(\omega)$ *are in* $L_2[0, \infty)$

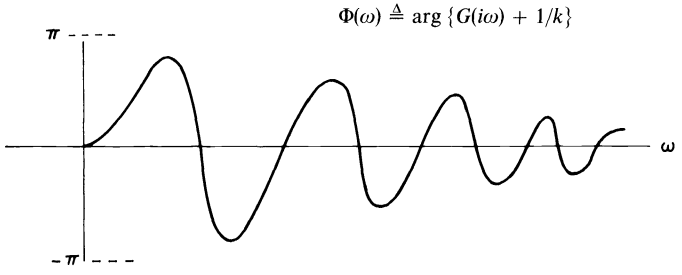


FIG. 2. *Plot of the phase function* $\Phi(\omega)$

with

$$\left(\int_0^\infty s(\omega)^2\,d\omega\right)^{1/2}\left(\int_0^\infty s'(\omega)^2\,d\omega\right)^{1/2} < \frac{(\log 2)^2}{2}. \tag{6.1}$$

Then $e(\cdot)$ *is in* $L_2[0,\infty)$ *and, in fact,* $\|e(\cdot)\| \leqq \text{const.}\,\|x(\cdot)\|$, *i.e., the feedback system pictured in Fig.* 1 *and described by* (2.1) (*or* (2.2)) *is* L_2*-stable.*

Proof. We may extend $s(\omega)$ to be defined on $(-\infty,\infty)$ by setting $s(\omega) = -s(-\omega)$ for $\omega < 0$. Then this extended $s(\omega)$ satisfies the conditions of Theorem 5.1 (in particular inequality (5.1) holds). Denoting this extended $s(\omega)$ again by $s(\omega)$, we see that (a) and (b) of this theorem hold now for $-\infty < \omega < \infty$.

Let $\mathcal{Z} \in \mathscr{L}$ given by $Z = \{1, y(\cdot)\}$ be the causal multiplier with $\arg Z(i\omega) = s(\omega)$ guaranteed by Theorem 4.1. Theorem 5.1 now assures that $\|y(\cdot)\|_1 < 1$.

In order to employ Theorem 3.5 to conclude L_2-stability, we now are left to show that $\text{Re}\,\{Z(i\omega)[G(i\omega) + 1/k]\} \geqq \delta$ for all real ω and suitable $\delta < 0$.

But both $|Z(i\omega)|$ and $|G(i\omega) + 1/k|$ attain a nonzero minimum over $-\infty < \omega < \infty$, as each is nonzero for finite ω and has a limit as $|\omega| \to \infty$. Therefore, there exists $\eta > 0$ with

$$|Z(i\omega)[G(i\omega) + 1/k]| \geqq \eta \qquad \text{for } -\infty < \omega < \infty. \tag{6.2}$$

Also,

$$\begin{aligned}\arg\{Z(i\omega)[G(i\omega) + 1/k]\} &= \arg\{Z(i\omega)\} + \arg\{G(i\omega) + 1/k\} \\ &= s(\omega) + \Phi(\omega),\end{aligned}$$

and so

$$|\arg\{Z(i\omega)[G(i\omega) + 1/k]\}| \leqq \rho < \pi/2 \tag{6.3}$$

for some constant ρ. This follows from (b) and the fact that both $s(\omega)$ and $\Phi(\omega)$ converge to zero as $|\omega| \to \infty$.

From (6.2) and (6.3) it now follows that

$$\text{Re}\,\{Z(i\omega)[G(i\omega) + 1/k]\} \geqq \eta \cos\rho > 0.$$

With this last verification, we are assured that Theorem 3.5 applies and the L_2-stability of feedback system (2.1) is proven.

Remark 6.2. It is not immediately obvious, but if conditions (a), (b) and (c) of Theorem 6.1 holds, then actually $-\pi/2 < s(\omega) < \pi/2$. For

$$\mathcal{Z} = \{1, y(\cdot)\} \quad \text{with} \quad \|y\|_1 < 1$$

implies $\text{Re}\,Z(i\omega) \geqq \varepsilon > 0$ for some $\varepsilon > 0$ and all ω, $-\infty < \omega < \infty$.

But this last condition implies $|\arg\{Z(i\omega)\}| < \pi/2$.

Remark 6.3. If condition (b) of Theorem 6.1 is replaced by

(b′) $|s(\omega) + \Phi(\omega)| < \pi/2$ for all $\omega > 0$,

the validity of the theorem remains. That this is true follows from noting that one may perturb $s(\omega)$ slightly so as to obtain the strict inequality (b) while maintaining the inequalities (a) and (c). For this perturbed $s(\omega)$, Theorem 6.1 will then apply as stated.

7. Applications. By making particular choices of $s(\omega)$, we may now employ Theorem 6.1 to yield specific stability criteria. In doing so it is useful to reorganize the results of § 6 somewhat by introducing the auxiliary "function"

$$\Psi(\omega) = \begin{cases} \pi/2 - \Phi(\omega), & \text{where } \Phi(\omega) \geqq 0, \\ -\pi/2 - \Phi(\omega), & \text{where } \Phi(\omega) \leqq 0. \end{cases}$$

Remark 7.1. (i) Note that where $\Phi(\omega) = 0$, $\Psi(\omega)$ is double-valued, i.e., $\Psi(\omega) = \pm\pi/2$. (ii) Condition (b′) of Remark 6.3 will be satisfied for any $s(\omega)$ lying within, or on the boundary of the unshaded region of Fig. 3. It is clear therefore that in order to apply Theorem 6.1, the object is to select an $s(\omega)$ (lying within the required region) satisfying $s(0) = 0$, $s(\omega)$ a.e. differentiable and with (6.1) holding.

DEFINITION 7.2. Let $\{I_j\}_{j=1,\cdots,n}$ be a maximal disjoint family of closed subintervals of $[0, \infty)$ (each containing at least two points) such that at the endpoints of each I_j, $|\Phi(\omega)| = \pi/2$, while $|\Phi(\omega)| \geqq \pi/2$ for all $\omega \in I_j$. See Fig. 4. (Note that since $\Phi(\omega) \to 0$ as $\omega \to \infty$, there can only be the finite number n of these I_j's.)

We are now ready to state the following theorem.

THEOREM 7.3. *Consider feedback equation* (2.1) *and assume the related Assumptions* 1–4 *hold. Let* $k > 0, \varepsilon > 0$ *be given and assume as in Theorem* 6.1 *that*

(i) $0 \leqq (N(r) - N(s))/(r - s) \leqq k - \varepsilon$ *for all real* r, s *with* $r \neq s$,

(ii) $g(\cdot) \in W_k$.

Further assume that $\Phi(\omega) \triangleq \arg\{G(i\omega) + 1/k\}$ *is a.e. differentiable. Then feedback system* (2.1) *will be* L_2*-stable provided that*

$$\left(\sum_{j=1}^{n} \int_{I_j} \Psi(\omega)^2 \, d\omega\right)\left(\sum_{j=1}^{n} \int_{I_j} \Phi'(\omega)^2 \, d\omega\right) < \frac{(\log 2)^4}{4}. \tag{7.1}$$

Proof. For $\omega \geqq 0$ define

$$s(\omega) = \begin{cases} \Psi(\omega), & \omega \in I_j, \text{ any } j, \\ 0, & \omega \notin I_j, \text{ any } j. \end{cases}$$

(See Fig. 4.)

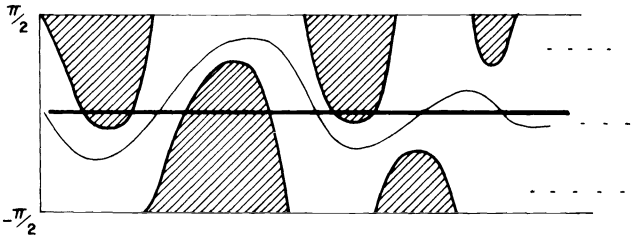


FIG. 3. *Plot of* $\Psi(\omega)$*—picture of an "acceptable"* $s(\omega)$

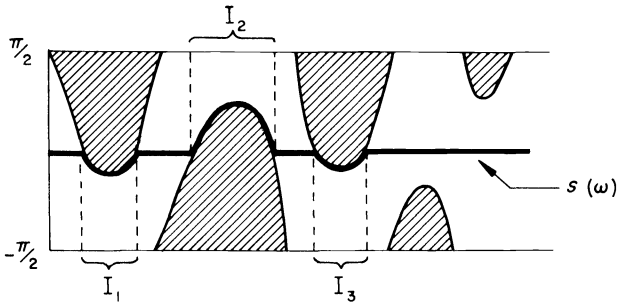


FIG. 4. *The $s(\omega)$ of Theorem* 7.3

Then all the conditions of Theorem 6.1 hold for this $s(\omega)$ except condition (b). In fact,

$$|s(\omega) + \Phi(\omega)| = \pi/2 \qquad \text{for } \omega \in I_j,$$

while

$$|s(\omega) + \Phi(\omega)| \leqq \pi/2 \qquad \text{for } \omega \notin I_j.$$

The weakened condition (b′) of Remark 6.3 therefore holds for all $\omega \geqq 0$. It follows, as pointed out in Remark 6.3, that system (2.1) is L_2-stable.

Remark 7.4.

(a) The assumption that $\Phi(\omega) = \arg\{G(i\omega) + 1/k\}$ is a.e. differentiable is easily met in practice. For instance, if $e^{\sigma t}g(t) \in L_1[0, \infty)$ for any $\sigma > 0$, then $G(s)$ will be analytic in $\operatorname{Re} s > -\sigma$ and, in this case, it easily follows that $\Phi'(\omega)$ exists everywhere (and is even continuous).

(b) Note that if $|\Phi(\omega)| < \pi/2$ for all $\omega \geqq 0$, then $\Psi(\omega) \equiv 0$ and so (7.1) holds trivially. This is the case where $\operatorname{Re}\{G(i\omega) + 1/k\} \geqq \delta > 0$, and has been much studied by Popov and others. The result is of course well known.

THEOREM 7.5. *Let the assumptions of Theorem* 7.3 *all hold except for inequality* (7.1). *Assume additionally that $\Phi(\omega)$ is continuously differentiable. Also let*

$$M_j \triangleq \max_{\omega \in I_j} |\Psi(\omega)|,$$

$$R_j \triangleq \max_{\omega \in I_j} |\Phi'(\omega)|,$$

and let $L_j \triangleq$ the length of I_j (note that by necessity $L_j \geqq 2M_j/R_j$ from their definitions).

Then the feedback system (2.1) *will be L_2-stable provided*

$$\sum_{j=1}^{n} M_j^2\left(L_j - \frac{4M_j}{3R_j}\right) + 2\sum_{j=1}^{n} R_j M_j < \frac{(\log 2)^4}{4}. \tag{7.2}$$

Proof. Let ω_j be the left-hand endpoint of I_j for $j = 1, \cdots, n$.

Define a function $s(\omega)$ for $\omega \geqq 0$ as follows:

(i) For $\omega \notin I_j$, any j, $s(\omega) = 0$.

(ii) For $\omega \in I_j$, if $\Psi(\omega) \geqq 0$ on I_j, let

$$s(\omega) = \begin{cases} R_j(\omega - \omega_j), & 0 \leqq \omega - \omega_j \leqq \dfrac{M_j}{R_j}, \\ M_j, & \dfrac{M_j}{R_j} \leqq \omega - \omega_j \leqq L_j - \dfrac{M_j}{R_j}, \\ R_jL_j - R_j(\omega - \omega_j), & L_j - \dfrac{M_j}{R_j} \leqq \omega - \omega_j \leqq L_j, \end{cases}$$

while if $\Psi(\omega) \leqq 0$ on I_j switch the sign of $s(\omega)$ as given above. (See Fig. 5.)

One can check that $s(\omega)$ as defined is an a.e. differentiable function (actually it is a polygonal path) whose graph lies in the "acceptable" region of the $\Psi(\omega)$ plot (see Fig. 5). Also inequality (7.2) is entirely equivalent to the inequality

$$\left(\int_0^\infty s(\omega)^2\, d\omega\right)\left(\int_0^\infty s'(\omega)^2\, d\omega\right) < (\log 2)^4$$

for this path. Hence Theorem 6.1 applies and L_2-stability is assured in this case.

8. Conclusion. The recipe developed in this paper for obtaining stability results is as follows:

Find an $s(\omega)$ which lies in the "acceptable" region of Fig. 3 and impose on it the inequality

$$\left(\int_0^\infty s(\omega)^2\, d\omega\right)\left(\int_0^\infty s'(\omega)^2\, d\omega\right) < \frac{(\log 2)^4}{4}. \tag{8.1}$$

Any such selection for $s(\omega)$ will yield a stability result.

By making specific choices for $s(\omega)$, one obtains particular results. One can, in fact, attempt to lend to these results predetermined chatacteristics by appropriately selecting $s(\omega)$.

For instance, in Theorem 7.3, the $s(\omega)$ is chosen to coincide with $\Psi(\omega)$ for certain ω-intervals and so the stability result in that instance involves an integral estimate of $|\Psi(\omega)|$ and $|\Psi'(\omega)| = |\phi'(\omega)|$. In Theorem 7.5, the choice of a polygonal $s(\omega)$ forces a different result, this time dependent on only the maximums of $\Psi(\omega)$ and $\phi'(\omega)$ over appropriate intervals.

One could go on. In general, it might seem most desirable to attempt to minimize the left side of (8.1) over all "acceptable" $s(\omega)$ in order to obtain a "best

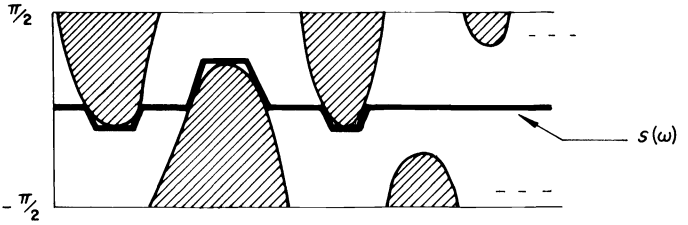


FIG. 5. *The polygonal $s(\omega)$ of Theorem 7.5*

possible" stability result. Unfortunately the minimization problem so posed does not appear to yield to analytic solution.

Final Remark 8.1. In much of the stability work by other authors, results are of a purely geometric nature, i.e., only the Nyquist plot

$$N_k = \{G(i\omega) + 1/k| \; - \infty < \omega < \infty\}$$

and its position with respect to a particular geometric set (circle, line, etc.) come into play.

In this paper, as in [4], this author has found that the phase function $\Phi(\omega)$ and its derivative $\Phi'(\omega)$ come into play. The presence of $\Phi'(\omega)$ in these results goes beyond purely geometric considerations. For instance, $G(i\omega)$ and $G_1(i\omega) = G(2i\omega)$ will have Nyquist plots which are geometrically identical. However, the different magnitudes of Φ'_1 and Φ' will indicate the different parametrizations. This author believes that the rate at which the Nyquist plot is swept out is strongly related to stability questions and feels that this present paper as well as [4] serve to indicate this.

REFERENCES

[1] R. P. O'SHEA, *A combined frequency-time domain stability criterion for autonomous continuous systems*, Proc. Joint Automatic Control Conference, University of Washington, Seattle, 1966, pp. 832–840.

[2] P. L. FALB AND G. ZAMES, *Stability conditions for systems with monotone and odd monotone non-linearities*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 221–223.

[3] M. I. FREEDMAN AND G. ZAMES, *Logarithmic variation criteria for the stability of systems with time-varying gains*, this Journal, 6 (1968), pp. 487–507.

[4] M. I. FREEDMAN, *L_2-stability of time-varying systems—construction of multipliers with prescribed phase characteristics*, this Journal, 6 (1960), pp. 559–578.

[5] R. W. BROCKETT AND L. J. FORYS, *On the stability of systems containing a time-varying gain*, Proc. Second Allerton Conference on Circuit and Systems Theory, Univ. of Illinois, Urbana, 1964, pp. 413–430.

[6] G. ZAMES, *On the input-output stability of time-varying non-linear feedback systems: I, II*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238; 465–476.

[7] C. A. DESOER, *A general formulation of the nyquist criterion*, IEEE Trans. Circuit Theory, CT-12 (1965), pp. 230–234.

[8] E. W. HOBSON, *The Theory of Functions of a Real Variable*, vol. 1, Dover, New York, 1957.

[9] E. C. TITCHMARSH, *Theory of Fourier Integrals*, 2nd. ed., Clarendon Press, Oxford, 1962.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

GLOBAL CONTROLLABILITY OF NONLINEAR SYSTEMS*

D. L. LUKES†

Abstract. This article deals with the controllability of nonlinear differential systems which arise when a linear system is perturbed. The sufficiency conditions presented insure that the nonlinear system will be (globally) controllable whenever its linear part is controllable. Moreover, the steering can be accomplished using continuous controls with arbitrarily prescribed initial and final values.

The problem of splitting the nonlinear system into controllable and uncontrollable parts is also discussed along with the nature of the null domain of controllability.

1. Introduction. Controllability theory attempts to define and isolate the theoretical limits to which a system can be controlled. Researchers are aware of its fundamental interconnections with other aspects of control including the existence of optimal controls, their feedback synthesis, stabilization and observability. An excellent account of these and other related topics can be found in the text [4] by Lee and Markus.

Our knowledge of the controllability of linear systems is certainly substantial. However, of the limited number of results in the nonlinear theory, most are in some sense local. Notable exceptions [1]–[3] study nonlinear systems in which the controls enter linearly. By restricting our attention to systems with a controllable linear part, we are able to obtain global results for systems in which the control can enter in a nonlinear fashion. The results cover linear systems as a simple special case and moreover show that the steering can be accomplished using continuous controls with arbitrarily prescribed initial and final values.

The differential control system to be considered is of the type

$$\dot{x} = f(u, x, t) = Ax + Bu + h(u, x, t) \tag{1.1}$$

with the standing assumption that the perturbation $h(u, x, t)$ is continuous at all $(u, x, t) \in R^{r+n+1}$. The linear part is given in terms of the real matrices A, B of respective sizes $n \times n$, $n \times r$.

The aim of the next section is to find conditions upon A, B and $h(u, x, t)$ which insure that for each $t_0, t_1 \in R^1$; $\alpha, \beta \in R^r$ and $x_0, x_1 \in R^n$ there exists a continuous control $u(\cdot):[t_0, t_1] \to R^r$ for (1.1) with $u(t_0) = \alpha$, $u(t_1) = \beta$ which produces a response $x(t)$ satisfying the boundary conditions $x(t_0) = x_0$, $x(t_1) = x_1$.

To simplify the notation, we make the preliminary translation of the origin $\tau = t - t_0$ which sends $t_0 \to 0$ and $t_1 \to T = t_1 - t_0$. This does not destroy the generality of the results.

2. Sufficient conditions for global controllability. The following observation will be vital to the perturbation analysis to follow.

* Received by the editors February 8, 1971, and in revised form June 2, 1971.

† Department of Applied Mathematics and Computer Science, University of Virginia, Charlottesville, Virginia 22901.

LEMMA 2.1. *Define* $\phi(\omega) = \int_0^\omega e^{A\sigma}B\,d\sigma$ *and let* $T > 0$. *If* rank $[B, AB, \cdots, A^{n-1}B] = n$, *then the matrix*

$$S_T = \int_0^T \phi(\omega)\phi^*(\omega)\,d\omega - \frac{1}{T}\left(\int_0^T \phi(\omega)\,d\omega\right)\left(\int_0^T \phi^*(\omega)\,d\omega\right) \tag{2.1}$$

is symmetric and positive definite.[1]

Proof. For nonzero $x \in R^n$ we compute the quadratic form

$$\begin{aligned} x \cdot S_T x &= \int_0^T |\phi^*(\omega)x|^2\,d\omega - \frac{1}{T}\left|\int_0^T \phi^*(\omega)x\,d\omega\right|^2 \\ &\geqq \int_0^T |\phi^*(\omega)x|^2\,d\omega - \frac{1}{T}\left(\int_0^T |\phi^*(\omega)x|\,d\omega\right)^2 \\ &\geqq \int_0^T |\phi^*(\omega)x|^2\,d\omega - \frac{1}{T}\int_0^T 1\,d\omega \int_0^T |\phi^*(\omega)x|^2\,d\omega = 0 \end{aligned} \tag{2.2}$$

by Schwarz's inequality in which equality holds only if $|\phi^*(\omega)x| = \lambda \cdot 1$ for all $\omega \in [0, T]$ with λ a constant. But clearly that constant would be zero and hence we would have $\phi^*(\omega)x = 0$ for $0 \leqq \omega \leqq T$. Repeated differentiation of the latter equation at $\omega = 0$ gives $B^*x = 0$, $B^*A^*x = 0, \cdots, B^*A^{*n-1}x = 0$. But $x \neq 0$ and we have the contradiction rank $[B, AB, \cdots, A^{n-1}B] < n$. Therefore $x \cdot S_T x > 0$ for all $x \neq 0$. Moreover it is trivial to check $S_T^* = S_T$.

The following notations will prove convenient:

$$C_T(t) = \int_{T-t}^T \phi^*(\omega)\,d\omega - \frac{t}{T}\int_0^T \phi^*(\omega)\,d\omega, \tag{2.3}$$

$$S_T(t) = \int_0^t e^{A(t-\tau)}BC_T(\tau)\,d\tau. \tag{2.4}$$

(We shall show $S_T(T) = S_T$.) Lemma 2.1 states a condition under which S_T is nonsingular and hence we write down the following system of equations whose importance will become apparent in the next lemma:

$$\begin{aligned} x(t) &= e^{At}x_0 + \phi(t)u_0 + \frac{1}{T}\int_0^t \phi(\omega)\,d\omega(u_T - u_0) \\ &\quad + S_T(t)y + \int_0^t e^{A(t-\omega)}h(u(\omega), x(\omega), \omega)\,d\omega, \end{aligned} \tag{2.5}$$

$$u(t) = \left(1 - \frac{t}{T}\right)u_0 + \left(\frac{t}{T}\right)u_T + C_T(t)y, \tag{2.6}$$

$$\begin{aligned} y = S_T^{-1}&\left[x_T - e^{AT}x_0 - \phi(T)u_0 - \frac{1}{T}\int_0^T \phi(\omega)\,d\omega(u_T - u_0)\right] \\ &- S_T^{-1}\int_0^T e^{A(T-\omega)}h(u(\omega), x(\omega), \omega)\,d\omega. \end{aligned} \tag{2.7}$$

[1] We denote the transpose of a matrix ϕ by ϕ^*.

LEMMA 2.2. *Any solution $x(t), u(t)$ to the nonlinear functional equations* (2.5)–(2.7) *provides a solution to the boundary value problem*

$$\dot{x}(t) = Ax(t) + Bu(t) + h(u(t), x(t), t), \tag{2.8}$$

$$x(0) = x_0, \qquad x(T) = x_T, \tag{2.9}$$

$$u(0) = u_0, \qquad u(T) = u_T. \tag{2.10}$$

Proof. Let $x(t)$, $u(t)$ be a solution to (2.5)–(2.7). Since $\phi(0) = 0$ and $S_T(0) = 0$, we see that $x(0) = x_0$. By noting that $C_T(0) = C_T(T) = 0$, we conclude from (2.6) that $u(0) = u_0$ and $u(T) = u_T$ as required.

The following computation shows that $S_T(T) = S_T$:

$$\begin{aligned} S_T(T) &= \int_0^T e^{A(T-\tau)} B C_T(\tau)\, d\tau \\ &= \int_0^T e^{A(T-\tau)} B \left[\int_{T-\tau}^T \phi^*(\omega)\, d\omega - \frac{\tau}{T} \int_0^T \phi^*(\omega)\, d\omega \right] d\tau \\ &= \int_0^T e^{A(T-\tau)} B \int_{T-\tau}^T \phi^*(\omega)\, d\omega\, d\tau - \frac{1}{T} \left(\int_0^T \phi(\omega)\, d\omega \right) \left(\int_0^T \phi^*(\omega)\, d\omega \right). \end{aligned} \tag{2.11}$$

Interchanging the order of integration, we have

$$\begin{aligned} \int_0^T e^{A(T-\tau)} B \int_{T-\tau}^T \phi^*(\omega)\, d\omega\, d\tau &= \int_0^T \int_{T-\omega}^T e^{A(T-\tau)} B\, d\tau\, \phi^*(\omega)\, d\omega \\ &= \int_0^T \int_0^\omega e^{A\sigma} B\, d\sigma\, \phi^*(\omega)\, d\omega \\ &= \int_0^T \phi(\omega) \phi^*(\omega)\, d\omega. \end{aligned} \tag{2.12}$$

Substitution of (2.12) into (2.11) and application of definition (2.1) shows $S_T(T) = S_T$. With this fact we can now compute, from (2.5) and (2.7),

$$\begin{aligned} x(T) = {} & e^{AT} x_0 + \phi(T) u_0 + \frac{1}{T} \int_0^T \phi(\omega)\, d\omega (u_T - u_0) \\ & + I \left[x_T - e^{AT} x_0 - \phi(T) u_0 - \frac{1}{T} \int_0^T \phi(\omega)\, d\omega (u_T - u_0) \right] \\ & - \int_0^T e^{A(T-\omega)} h(u(\omega), x(\omega), \omega)\, d\omega + \int_0^T e^{A(T-\omega)} h(u(\omega), x(\omega), \omega)\, d\omega \\ = {} & x_T. \end{aligned} \tag{2.13}$$

Therefore all the boundary conditions are satisfied. All that remains to be shown is that $x(t)$, $u(t)$ satisfy (2.8).

Using (2.6), we first compute

$$\begin{aligned}\int_0^t e^{A(t-\tau)}Bu(\tau)\,d\tau &= \int_0^t e^{A(t-\tau)}B\left[\left(1-\frac{\tau}{T}\right)u_0+\left(\frac{\tau}{T}\right)u_T+C_T(\tau)y\right]d\tau\\ &= \int_0^t e^{A(t-\tau)}B\,d\tau\,u_0+\frac{1}{T}\int_0^t e^{A(t-\tau)}\tau B\,d\tau(u_T-u_0)\\ &\quad+\int_0^t e^{A(t-\tau)}BC_T(\tau)\,d\tau\,y\\ &= \phi(t)u_0+\frac{1}{T}\int_0^t \phi(\omega)\,d\omega(u_T-u_0)+S_T(t)y.\end{aligned} \tag{2.14}$$

Combining this equation with (2.5) shows that

$$x(t)-\left[e^{At}x_0+\int_0^t e^{A(t-\tau)}Bu(\tau)\,d\tau+\int_0^t e^{A(t-\tau)}h(u(\tau),x(\tau),\tau)\,d\tau\right]=0. \tag{2.15}$$

But this equation can be differentiated to give

$$\dot{x}(t)-[Ax(t)+Bu(t)+h(u(t),x(t),t)]=0 \tag{2.16}$$

and the proof is complete.

Remark 2.1. Any solution $x(t)$, $u(t)$ to (2.5)–(2.7) will have $u(t)$ analytic in t and $x(t)$ continuous along with its first derivative. In general, the differentiability will be limited only by the differentiability of h.

The following theorem presents our first sufficiency conditions which insure the global controllability of (1.1).

THEOREM 2.1. *Consider the differential control system in* R^n,

$$\dot{x}=f(u,x,t)=Ax+Bu+h(u,x,t), \tag{2.17}$$

with $h(u,x,t)$ *continuous and* $|h(u,x,t)|$ *bounded on* R^{r+n+1}. *Further assume that* $h(u,x,t)$ *is periodic in* t *with period* $T>0$ *for each fixed* u, x. *Finally assume that* rank $[B, AB, \cdots, A^{n-1}B]=n$.

Then for each α, $\beta\in R^r$; $x_0, x_1\in R^n$ *and* $0<t_1<T$ *there exists a control* $u(\cdot)\in C^\omega([0,T],R^r)$ *such that*

(i) $u(0)=u(T)=\alpha$, $u(t_1)=\beta$ *and*

(ii) *a corresponding solution of* (2.17) *for which* $x(0)=x_0$ *satisfies both* $x(t_1)=x_1$, $x(T)=x_0$.

This control extends to a continuous periodic function on $(-\infty,\infty)$ *with period* T *and its response agrees with the periodic extension of* $x(t)$.

Remark 2.2. We do not require that the period T be a minimal period. Hence the theorem applies to the special case where $h=h(u,x)$ is independent of t (autonomous systems). In particular, it applies to the case where $h\equiv 0$ (linear homogeneous systems). But even in the latter case it still says more than what is usually stated (see [4, p. 32])—namely, not only can the response be steered through any two states but moreover with any attainable prescribed velocities.

Remark 2.3. We may drop the periodicity requirement on $h(u, x, t)$ altogether. Then $T > 0$ is arbitrary. The conclusions of the theorem are all valid with the exception of the last sentence concerning the periodicity which is lost.

Proof of Theorem 2.1. The proof will be based upon two applications of Lemma 2.2. In that lemma we first let the role of T be played by t_1. Hence we wish to establish the existence of a solution $x(t)$, $u(t)$ to (2.5)–(2.7) with $u_0 = \alpha$, $u_{t_1} = \beta$, $x(0) = x_0$ and $x(t_1) = x_1$. For this purpose, consider the sequence of successive approximations on $0 \leqq t \leqq t_1$:

$$x^{(k+1)}(t) = e^{At}x_0 + \phi(t)\alpha + \frac{1}{t_1}\int_0^t \phi(\omega)\,d\omega(\beta - \alpha) + S_{t_1}(t)y^{(k)} + \int_0^t e^{A(t-\omega)}h(u^{(k)}(\omega), x^{(k)}(\omega), \omega)\,d\omega, \tag{2.18}$$

$$u^{(k+1)}(t) = \left(1 - \frac{t}{t_1}\right)\alpha + \left(\frac{t}{t_1}\right)\beta + C_{t_1}(t)y^{(k)}, \tag{2.19}$$

$$y^{(k)} = S_{t_1}^{-1}\left[x_{t_1} - e^{At_1}x_0 - \phi(t_1)\alpha - \frac{1}{t_1}\int_0^{t_1}\phi(\omega)\,d\omega(\beta - \alpha)\right] - S_{t_1}^{-1}\int_0^{t_1} e^{A(t_1-\omega)}h(u^{(k)}(\omega), x^{(k)}(\omega), \omega)\,d\omega, \tag{2.20}$$

$k = 0, 1, 2, \cdots$, with, for example, $x^{(0)}(t) \equiv 0$, $u^{(0)}(t) \equiv 0$.

Let b_1 be the bound on $|h(u, x, t)|$. From (2.20) it is clear that $|y^{(k)}| \leqq b_2$, $k = 0, 1, 2, \cdots$, for some bound b_2 independent of k. With these bounds and the continuity and hence uniform continuity of the terms in (2.18) independent of k it follows directly that the sequence $\{x^{(k)}(t)\}$ is a uniformly bounded, equi-uniformly continuous sequence of functions on $0 \leqq t \leqq t_1$. By the Arzela–Ascoli theorem we conclude that there exists a subsequence converging uniformly on $[0, t_1]$ to a continuous limit function $x^\infty(t)$. A further subsequence can be extracted such that the corresponding $y^{(k)}$ converge since $|y^{(k)}| \leqq b_2$, $k = 0, 1, 2, \cdots$. Hence there is a subsequence $x^{(k_i)}(t) \to x^\infty(t)$ uniformly on $[0, t_1]$ with corresponding $y^{(k_i)} \to y^\infty$ as $i \to \infty$. But (2.19) shows that this forces $u^{(k_i)}(t) \to u^\infty(t)$ uniformly on $[0, t_1]$ as $i \to \infty$ for some continuous limit function $u^\infty(t)$. With the aid of the continuity of $h(u, x, t)$, it follows that the limits satisfy the system of equations obtained upon substitution of $x^\infty(t)$, $u^\infty(t)$ and y^∞ for the respective terms in k in (2.18)–(2.20). In this way, we have obtained a solution to the corresponding boundary value problem described in Lemma 2.2.

By our construction, $x^\infty(t_1) = x_1$ and $u^\infty(t_1) = \beta$. With a similar construction, using initial values x_1, β at t_1 and final values x_0, α at T our solution $x^\infty(t)$, $u^\infty(t)$ can be extended to $[0, T]$ as a solution to (2.17) satisfying all the boundary conditions: $x^\infty(0) = x_0 = x^\infty(T)$, $u^\infty(0) = \alpha = u^\infty(T)$, $x^\infty(t_1) = x_1$, $u^\infty(t_1) = \beta$. Furthermore, since $h(u, x, t)$ is periodic with period T, $f(u^\infty(0), x^\infty(0), 0) = f(\alpha, x_0, 0) = f(\alpha, x_0, T) = f(u^\infty(T), x^\infty(T), T)$ which now allows us to further extend $u^\infty(t)$ by periodicity to $(-\infty, \infty)$ and its response agrees with the periodic extension of $x^\infty(t)$ to $(-\infty, \infty)$.

Remark 2.4. We choose to avoid a formal definition of controllability. In the linear theory the usual definition as the capability to steer the response through any

two states in finite time turns out to be equivalent to the condition rank $[B, AB, \cdots, A^{n-1}B] = n$. Hence we have shown that the controllability of a linear system is preserved when we add a perturbation $h(u, x, t)$ with $|h(u, x, t)|$ bounded.

Now we turn our attention to a different condition upon $h(u, x, t)$ which allows $|h(u, x, t)|$ to be unbounded; namely,

$$|h(u, x, t) - h(v, y, t)| \leqq L[|u - v| + |x - y|] \tag{2.21}$$

for some constant L, holding for all $u, v \in R^r$; $x, y \in R^n$ and $t \in R^1$.

2.1. A preliminary estimate. Again we consider the successive approximations to (2.5)–(2.7):

$$\begin{aligned} x^{(k+1)}(t) &= e^{At}x_0 + \phi(t)u_0 + \frac{1}{T}\int_0^t \phi(\omega)\,d\omega(u_T - u_0) \\ &\quad + S_T(t)y^{(k)} + \int_0^t e^{A(t-\omega)}h(u^{(k)}(\omega), x^{(k)}(\omega), \omega)\,d\omega, \end{aligned} \tag{2.22}$$

$$u^{(k+1)}(t) = \left(1 - \frac{t}{T}\right)u_0 + \left(\frac{t}{T}\right)u_T + C_T(t)y^{(k)}, \tag{2.23}$$

$$\begin{aligned} y^{(k)} &= S_T^{-1}\left[x_T - e^{AT}x_0 - \phi(T)u_0 - \frac{1}{T}\int_0^T \phi(\omega)\,d\omega(u_T - u_0)\right] \\ &\quad - S_T^{-1}\int_0^T e^{A(T-\omega)}h(u^{(k)}(\omega), x^{(k)}(\omega), \omega)\,d\omega, \end{aligned} \tag{2.24}$$

$k = 0, 1, 2, \cdots$, with $x^{(0)}(t) \equiv 0$, $u^{(0)}(t) \equiv 0$.

Before estimating this sequence we make a few observations. If one introduces a change of control variables $u = \tilde{u} + Dx$ in (1.1), where D is a matrix, then A transforms into $A + BD$ and (2.21) is preserved (with proper readjustment of L). Since rank $[B, AB, \cdots, A^{n-1}B] = n$ is assumed, it follows that D can be chosen so that $A + BD$ is a stability matrix (see [5, p. 45] or [6]) and moreover the rank condition is preserved. Therefore with no loss of generality we assume that A is already a stability matrix and thus have available the estimate

$$|e^{At}| = |e^{A^*t}| \leqq c\,e^{-\mu t}, \qquad t \geqq 0, \tag{2.25}$$

for a constant c depending upon A and μ a positive number such that $\operatorname{Re}\lambda < -\mu$ for each eigenvalue λ of A. Noting that $\det(A) \neq 0$, we also have the estimate

$$|A^{-1}| = \left|-\int_0^\infty e^{At}\,dt\right| \leqq c\int_0^\infty e^{-\mu t}\,dt = \frac{c}{\mu}. \tag{2.26}$$

With the aid of (2.21), (2.25) and (2.26), we proceed to estimate the differences computed from (2.22)–(2.24):

$$\begin{aligned} (x^{k+1} - x^k)(t) &= S_T(t)(y^k - y^{k-1}) + \int_0^t e^{A(t-\omega)}[h^k - h^{k-1}]\,d\omega \\ &= -S_T(t)S_T^{-1}\int_0^T e^{A(T-\omega)}[h^k - h^{k-1}]\,d\omega \\ &\quad + \int_0^t e^{A(t-\omega)}[h^k - h^{k-1}]\,d\omega, \end{aligned} \tag{2.27}$$

$$|(x^{k+1} - x^k)(t)| \leqq |S_T(t)S_T^{-1}| \int_0^T |e^{A(T-\omega)}|\, d\omega L[\|u^k - u^{k-1}\| + \|x^k - x^{k+1}\|]$$

$$+ \int_0^t |e^{A(t-\omega)}|\, d\omega L[\|u^k - u^{k-1}\| + \|x^k - x^{k-1}\|], \tag{2.28}$$

$$|(u^{k+1} - u^k)(t)| \leqq |C_T(t)(y^k - y^{k-1})|$$

$$\leqq |C_T(t)S_T^{-1}| \int_0^T |e^{A(T-\omega)}|\, d\omega L[\|u^k - u^{k-1}\| + \|x^k - x^{k-1}\|]. \tag{2.29}$$ [2]

By estimating the integrals in (2.28)–(2.29) and adding, one gets

$$\|x^{k+1} - x^k\| + \|u^{k+1} - u^k\| \leqq \eta_T[\|x^k - x^{k-1}\| + \|u^k - u^{k-1}\|], \tag{2.30}$$

where

$$\eta_T = (c/\mu)L[1 + \|S_T(\cdot)S_T^{-1}\| + \|C_T(\cdot)S_T^{-1}\|]. \tag{2.31}$$

THEOREM 2.2. *Consider the one-parameter family of differential control systems in* R^n :

$$\dot{x} = f_\varepsilon(u, x, t) = Ax + Bu + \varepsilon h(u, x, t) \tag{2.32}$$

with $h(u, x, t)$ *continuous and*

$$|h(u, x, t) - h(v, y, t)| \leqq L[|u - v| + |x - y|] \tag{2.33}$$

on R^{r+n+1} *for a constant* L. *Further, assume that* $h(u, x, t)$ *is periodic in* t *with period* $T > 0$ *for each fixed* u, x. *Suppose that* rank $[B, AB, \cdots, A^{n-1}B] = n$.

Then the conclusion in Theorem 2.1 *and the remarks following it are valid for* (2.32) *for all* ε *in a neighborhood of* $\varepsilon = 0$.

Proof. Consider the corresponding sequences $x^{(k)}(t)$, $u^{(k)}(t)$ and $y^{(k)}$ which arise in the proof of Theorem 2.1. The estimate (2.30) is valid with η_T replaced by $|\varepsilon|\eta_T$. By choosing $|\varepsilon|$ small so that $|\varepsilon|\eta_T < 1$, we are assured $x^{(k)}(t)$, $u^{(k)}(t)$ converge uniformly. The remainder of the proof proceeds exactly as in Theorem 2.1.

2.2. Estimation of $C_T(t)$, $S_T(t)$ and S_T^{-1}. The next theorem discusses the controllability of (2.1) wherein the orbit time T is not prescribed a priori but dictated by (2.1) itself. The achievement of this result requires a detailed estimate of η_T defined in (2.31) which we proceed to establish.

In view of our assumption that rank $[B, AB, \cdots, A^{n-1}B] = n$, we suppose the previously discussed change of control variable has been made so that A is a stability matrix and the estimates (2.25), (2.26) and (2.30) are all available. Since A is then nonsingular, we can compute

$$\phi(\omega) = \int_0^\omega e^{A\sigma}B\, d\sigma = (e^{A\omega} - I)A^{-1}B. \tag{2.34}$$

[2] We use the matrix norm $|M| = \sup_{|x|=1} |Mx|$ and notations $\|u\| = \sup_{[0,T]} |u(t)|$, etc.

Starting with the defining formulas (2.3) and (2.4), one can further compute

$$C_T(t) = B^*A^{*-2}\left[\left(1 - \frac{t}{T}\right)e^{A^*T} - \left(e^{A^*(T-t)} - \frac{t}{T}\right)\right], \tag{2.35}$$

$$\begin{aligned} S_T(t) = {} & \left(A^{-1}\int_0^t e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma A^{*-1}\right)e^{A^*(T-t)} \\ & - \frac{1}{T}(e^{At} - I)A^{-2}BB^*A^{*-2}(e^{A^*T} - I) - A^{-1}BC_T(t), \end{aligned} \tag{2.36}$$

$$\begin{aligned} S_T(T) = {} & A^{-1}\int_0^T e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma A^{*-1} \\ & - \frac{1}{T}(e^{AT} - I)A^{-2}BB^*A^{*-2}(e^{A^*T} - I). \end{aligned} \tag{2.37}$$

By factoring out the term containing the integral in (2.37) and using the fact that

$$(e^{AT} - I) \quad \text{and} \quad \int_0^T e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma$$

converge as $T \to \infty$, one can easily show that

$$S_T^{-1}(T) = A^*\left(\int_0^T e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma\right)^{-1}A + O_1(T), \tag{2.38}$$

where $\|O_1(T)\| = O(1/T)$. Multiplying, we obtain

$$\begin{aligned} S_T(t)S_T^{-1}(T) = {} & A^{-1}\left[\left(\int_0^t e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma\right)e^{A^*(T-t)}\right. \\ & \left. + BB^*A^{*-1}\left(e^{A^*(T-t)} - \frac{t}{T}\right)\right]\left(\int_0^T e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma\right)^{-1}A \\ & + O_2(T), \end{aligned} \tag{2.39}$$

where $\|O_2(T)\| = O(1/T)$.

Integration by parts shows

$$\begin{aligned} \int_0^t e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma = {} & e^{At}BB^*\, e^{A^*t}A^{*-1} - BB^*A^{*-1} \\ & - A\int_0^t e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma A^{*-1} \end{aligned} \tag{2.40}$$

which can be substituted into (2.39) to give

$$\begin{aligned} S_T(t)S_T^{-1}(T) = {} & -\left[\left(\int_0^t e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma\right)e^{A^*(T-t)} + A^{-1}BB^*\frac{t}{T}\right] \\ & \cdot\left[A^{*-1}\left(\int_0^T e^{A\sigma}BB^*\, e^{A^*\sigma}\, d\sigma\right)^{-1}A\right] + O_4(T) \end{aligned} \tag{2.41}$$

with $\|O_4(T)\| = O(1/T)$.

This leads directly to the estimate

$$\|S_T(t)S_T^{-1}(T)\| \leqq \frac{3c^4|B|^2|A|}{2\mu^2}\left|\left(\int_0^T e^{A\sigma}BB^* e^{A^*\sigma}\,d\sigma\right)^{-1}\right| + O(1/T). \tag{2.42}$$

In a similar fashion we see that

$$C_T(t)S_T^{-1}(T) = -B^*A^{*-1}\left(e^{A^*(T-t)} - \frac{t}{T}\right)\left(\int_0^T e^{A\sigma}BB^* e^{A^*\sigma}\,d\sigma\right)^{-1} A + O_5(T). \tag{2.43}$$

If we set $1 - t/T = \sigma$ and

$$A^{*-1}\left(e^{A^*T(1-t/T)} - \frac{t}{T}\right) = A^{*-1}(e^{A^*T\sigma} + \sigma - 1) = \psi(\sigma), \tag{2.44}$$

then $\psi(0) = 0$ and $d\psi/d\sigma = (A^*T)\psi + T(1-\sigma) + A^{*-1}$.

Solving this differential equation gives

$$\psi(\sigma) = \int_0^\sigma e^{A^*T(\sigma-\omega)}[T(1-\omega) + A^{*-1}]\,d\omega \tag{2.45}$$

for $0 \leqq \sigma \leqq 1$. Estimation shows

$$\begin{aligned} |\psi(\sigma)| &\leqq \int_0^\sigma c\, e^{-\mu T(\sigma-\omega)}[T + |A^{*-1}|]\,d\omega \\ &= [T + |A^{*-1}|]\frac{c(1 - e^{-T\mu\sigma})}{\mu T} \leqq \frac{c}{\mu}\left(1 + \frac{|A^{*-1}|}{T}\right) \\ &= \frac{c}{\mu} + O(1/T). \end{aligned} \tag{2.46}$$

Application of (2.46) to (2.43) produces the estimate

$$\|C_T(\cdot)S_T^{-1}(T)\| \leqq \frac{c|B||A|}{\mu}\left|\left(\int_0^T e^{A\sigma}BB^* e^{A^*\sigma}\,d\sigma\right)^{-1}\right| + O(1/T). \tag{2.47}$$

For $0 < T \leqq \infty$ define λ_T to be the smallest eigenvalue of $\int_0^T e^{A\sigma}BB^* e^{A^*\sigma}\,d\sigma$. By the theory of symmetric matrices,

$$\left|\left(\int_0^T e^{A\sigma}BB^* e^{A^*\sigma}\,d\sigma\right)^{-1}\right| = \frac{1}{\lambda_T}. \tag{2.48}$$

From the definition (2.31), (2.42), (2.47) and (2.48) we conclude, after adding inequalities, that

$$\eta_T \leqq \frac{cL}{\mu}\left[1 + \frac{|A||B|c}{\mu\lambda_T}\left(1 + \frac{3|B|c^3}{2\mu}\right)\right] + O(1/T). \tag{2.49}$$

If we define

$$\eta_\infty = \frac{cL}{\mu}\left[1 + \frac{|A||B|c}{\mu\lambda_\infty}\left(1 + \frac{3|B|c^3}{2\mu}\right)\right], \tag{2.50}$$

we see that if $\eta_\infty < 1$, then $\eta_T < 1$ for all sufficiently large T.

Our results can now be summarized in the next theorem.

THEOREM 2.3. *Assert the hypothesis of Theorem* 2.1 *but replace the boundedness of* $|h(u, x, t)|$ *by* (2.21). *Recall A can be assumed to be a stability matrix with ensuing* (2.21), (2.25), (2.48) *and* (2.50) *defining* η_∞.

If $\eta_\infty < 1$, *then there exist a* t_1 *and an integer k satisfying* $0 < t_1 < kT$ *such that for each* $\alpha, \beta \in R^r$; $x_0, x_1 \in R^n$ *there exists a control* $u(\cdot) \in C^\omega([0, kT], R^r)$ *for which*

(i) $u(0) = u(kT) = \alpha$, $u(t_1) = \beta$ *and*

(ii) *the corresponding solution of* (2.17) *for which* $x(0) = x_0$ *satisfies both* $x(t_1) = x_1$, $x(kT) = x_0$.

This control extends to a continuous periodic function on $(-\infty, \infty)$ *with period kT and its response agrees with the periodic extension of* $x(t)$.

Remarks 2.2, 2.3 *concerning autonomous and nonperiodic systems remain valid.*

Proof. The proof proceeds as in Theorem 2.1 with two applications of Lemma 2.2. If $\eta_\infty < 1$, then, as noted after (2.50), $\eta_{t_1} < 1$ for all sufficiently large t_1. We let any solution of this inequality determine the required t_1 discussed in the statement of Theorem 2.3. Hence the approach is to first obtain that part of the control steering x_0 to x_1 on $0 \leqq t \leqq t_1$. It is determined as in Theorem 2.1 through the successive approximations (2.18)–(2.20) which converge uniformly since they satisfy an inequality (2.30) in which the parameter η_T is now $\eta_{t_1} \leqq \eta_\infty < 1$. The second part of the control is obtained in a similar fashion by first choosing k large so that $kT - t_1 \geqq t_1$ with new initial control value β, initial state x_1, etc.

Remark 2.5. Example 2.1 below points out the fact that the global controllability of a linear system can be destroyed by a nonlinear perturbation $h(u, x)$ satisfying (2.21). Thus the requirements that $|\varepsilon|$ be small in Theorem 2.2 and $\eta_\infty < 1$ in Theorem 2.3 are not superfluous. The example also demonstrates how η_∞ depends upon the choice of preliminary change of control variable $u \to u + Dx$ making A a stability matrix.

Example 2.1. Consider the scalar system

$$\dot{x} = f_\varepsilon(u, x) = u + \varepsilon\sqrt{u^2 + x^2}. \tag{2.51}$$

It is clear that $\varepsilon\dot{x}(t) \geqq 0$ for all $t \in R^1$ if $|\varepsilon| \geqq 1$. This shows that with ε in that range the conclusions of Theorems 2.2, 2.3 do not hold. Introduction of the change of control variable $u \to u - \lambda x$ with $\lambda > 0$ transforms (2.51) into the stabilized system

$$\dot{x} = -\lambda x + u + h(u, x) \tag{2.52}$$

in which $h(u, x) = \varepsilon((u - \lambda x)^2 + x^2)^{1/2}$. One may easily check that $h(u, x)$ satisfies (2.21) with $L = |\varepsilon|(1 + \lambda)$. We note that $\mu = \lambda$, $c = 1$ and from (2.48) compute $\lambda_\infty = 1/(2\lambda)$. Substitution of these parameters into (2.50) gives

$$\eta_\infty = |\varepsilon|(1 + 1/\lambda)[1 + 2\lambda(1 + 3/(2\lambda))]. \tag{2.53}$$

If $|\varepsilon| < .08$, then by choosing $\lambda = \sqrt{2}$ we can make $\eta_\infty < 1$ and Theorem 2.3 applies. One can easily show that global controllability persists for $|\varepsilon| < 1$.

By discarding the square root in (2.51) we obtain the equation

$$\dot{x} = f_\varepsilon(u, x) = u + \varepsilon(u^2 + x^2) \tag{2.54}$$

which fails to be globally controllable for all ε but $\varepsilon = 0$.

3. Partially controllable systems. This section discusses the controllability of (1.1) in the case

$$\text{rank } [B, AB, \cdots, A^{n-1}B] = d < n. \tag{3.1}$$

One can develop the well-known linear theory by establishing the existence of a maximal controllable linear subspace in R^n of dimension d (see, e.g., [4, p. 97]). This gives rise to a nonsingular real linear change of state coordinates which transforms (1.1) into the form

$$\dot{x}_1 = A_{11}x_1 + B_1u + A_{12}x_2 + h_1(u, x_1, x_2, t), \tag{3.2}$$

$$\dot{x}_2 = A_{22}x_2 + h_2(u, x_1, x_2, t) \tag{3.3}$$

with $\dim (x_1) = d$, $\dim (x_2) = n - d$ and

$$\text{rank } [B_1, A_{11}B_1, \cdots, A_{11}^{d-1}B_1] = d. \tag{3.4}$$

We assume $d \geqq 1$. The terms h_1 and h_2 represent h in the new coordinate system. Since (3.1) and most controllability properties of interest are preserved by nonsingular linear transformations, we confine our discussion of the controllability of (1.1) to the model (3.2)–(3.3) satisfying (3.4). In general, the behavior of such a system can be very complicated. However the results of § 2 do provide useful information in some special cases.

3.1. The special case $h_2(u, x_1, x_2, t) = h_2(x_2, t)$. If $h_2(u, x_1, x_2, t)$ turns out to be independent of both u and x_1, then (3.2)–(3.3) can be written as

$$\dot{x}_1 = A_{11}x_1 + B_1u + h_1(u, x_1, t), \tag{3.5}$$

$$\dot{x}_2 = A_{22}x_2 + h_2(x_2, t) \tag{3.6}$$

in which

$$h_1(u, x_1, t) = A_{12}x_2(t) + h_1(u, x_1, x_2(t), t) \tag{3.7}$$

and $x_2(t)$ is a solution of (3.6). The x_2-coordinates are called *uncontrollable* since they are independent of u and x_1. On the other hand, x_1 can be regarded as a *controllable* system of coordinates since (3.5) is again of the type (1.1) satisfying the full rank condition (3.4) treated in § 2. Note from (3.7) that even with (1.1) autonomous, the condition (3.1) precludes a treatment in terms of autonomous systems if A_{12} is not zero.

Of course, in studying the controllability of (3.5), we must keep in mind the fact that $h_1(u, x_1, t)$ depends upon $x_2(t)$, the particular solution of (3.6) chosen. To illustrate how Theorem 2.1 can be applied, suppose $|h(u, x, t)|$ is bounded. Then $|h_2(x_2, t)|$ will be bounded. Consequently (3.6) will have a solution $x_2(t)$ on every compact interval through each initial state $x_2(0)$. But this makes $|h_1(u, x_1, t)|$ computed from (3.7) bounded, and Theorem 2.1 now applies to (3.5).[3]

Similarly, if the original h satisfies (2.21), then (3.6) will have a solution on every compact interval through each initial state. It follows that $h_1(u, x_1, t)$ then satisfies a condition of type (2.21) and Theorem 3.1 then applies to (3.5). Note

[3] The proof of Theorem 2.1 only requires boundedness of $|h(u, x, t)|$ on $R^{r+n} \times [0, T]$.

that if $h_1(u, x_1, x_2, t)$ is independent of x_2, then the constant L can be taken independent of the solution of (3.6) selected in (3.7).

3.2. Other cases. Minor variations in the techniques of §2 allow one to treat other systems. Suppose, for example, that $|h(u, x, t)|$ were bounded so that $|h_1(u, x_1, x_2, t)|$ and $|h_2(u, x_1, x_2, t)|$ would likewise be bounded. In general, this in itself does not allow the response to be steered through arbitrary initial and final states. However, according to the next theorem, the controllable part of the state can be steered in such a fashion.

THEOREM 3.1. *Let $T > 0$ and consider the differential control system* (3.2)–(3.4) *in R^n in which $h_1(u, x_1, x_2, t)$, $h_2(u, x_1, x_2, t)$ are continuous and $|h_1(u, x_1, x_2, t)|$, $|h_2(u, x_1, x_2, t)|$ bounded on $R^{r+n} \times [0, T]$.*

Then for each u_0, $u_T \in R^r$; x_1^0, $x_1^T \in R^d$ and $x_2^0 \in R^{n-d}$ there exists a control $u(\cdot) \in C^\omega([0, T], R^r)$ such that

(i) *$u(0) = u_0$, $u(T) = u_T$ and*

(ii) *a corresponding solution of* (3.2)–(3.3) *satisfying $x_1(0) = x_1^0$, $x_2(0) = x_2^0$ also satisfies $x_1(T) = x_1^T$.*

Proof. The proof can be carried out by setting up a staggered sequence wherein the approximation $u^{(k-1)}(t)$ from (2.23) would be substituted into (3.3) along with $x_1{}^{(k-1)}(t)$ to produce $x_2{}^{(k-1)}(t)$ and h in (2.22) and (2.24) would be replaced by $h_1(u^{(k)}(\omega), x_1{}^{(k)}(\omega), x_2{}^{(k-1)}(\omega), \omega)$. The boundedness of $|h_1|$ and $|h_2|$ can then be applied to establish the uniform boundedness of $|x_2{}^{(k)}(t)|$, $|x_1{}^{(k)}(t)|$ and permits the extraction of the required uniformly convergent subsequences to complete the proof along the lines of Theorem 2.1.

Remark 3.1. A comparison of the following examples in R^2 shows that the extent to which (3.2)–(3.4) is controllable is heavily dependent upon the term $h_2(u, x_1, x_2, t)$.

Example 3.1.

$$\dot{x}_1 = u,$$

$$\dot{x}_2 = x_1 u^2.$$

Example 3.2.

$$\dot{x}_1 = u,$$

$$\dot{x}_2 = x_1 u.$$

Any initial state $x_1(0) = x_1^0$, $x_2(0) = x_2^0$ can be connected to any final state $x_1(T) = x_1^T$, $x_2(T) = x_2^T$ by the response to an appropriate control $u(t)$ in Example 3.1, whereas the responses of Example 3.2 are restricted to parabolas.

4. The domain C of null controllability. The previous sections were concerned with finding conditions upon A, B and $h(u, x, t)$ in (1.1) which assure that one can steer between any two prescribed states in finite time. Then x^0 can be steered to x^T and x^T can be steered to x^0. However, in applications where x denotes the error in some controlled physical system, the point of interest is the capability of steering all initial states to one fixed state, the origin. This leads to the definition of the *domain of null controllability* given below. Our discussion will be restricted to autonomous systems in R^n of the type

$$\dot{x} = Ax + Bu + h(u, x), \tag{4.1}$$

with the standing assumptions that h together with its first partial derivatives are continuous in R^{r+n} and zero at the origin in R^{r+n}.

DEFINITION 4.1. The *domain C of null controllability* for (4.1) consists of all initial states in R^n which can be steered to the origin in finite time using continuous controls.

As a corollary to the theorems of § 2, we have the following theorem.

THEOREM 4.1. *If for the control system* (4.1) *in* R^n,

$$\text{rank } [B, AB, \cdots, A^{n-1}B] = n, \tag{4.2}$$

and one of the conditions

(i) $|h(u, x)|$ *bounded on* R^{r+n},

(ii) $|h(u, x) - h(v, y)| \leqq L[|u - v| + |x - y|]$ *for all* $u, v \in R^r$; $x, y \in R^n$ *for a constant* L, *with* $\eta_\infty < 1$

holds, then the domain C of null controllability is R^n.

In [4, p. 366], Lee and Markus prove that if (4.2) holds, then C is an open subset of R^n containing the origin. Their definition of C does not require continuity of the controls but imposes the control constraint $u(t) \in \Omega \subseteq R^r$ where Ω is compact. Their hypothesis includes the assumption that $u = 0$ is interior to Ω.

The following local result amounts to a slight generalization of the cited theorem, in that (4.2) is relaxed. It is valid for either definition of C.

THEOREM 4.2. *If in the differential control system in* R^n,

$$\dot{x} = f(u, x) = Ax + Bu + h(u, x), \tag{4.3}$$

$$\text{rank } [B, AB, \cdots, A^{n-1}B] = d \leqq n, \tag{4.4}$$

then C contains a d-dimensional submanifold of R^n containing the origin.

Proof. We outline the proof, since it is quite similar to the one in [4, p. 366]. One considers the system

$$\dot{x} = -f(u, x) \tag{4.5}$$

and studies the collection of states which can be attained by steering from the origin as the initial state. This set is contained in C.

By an appropriate nonsingular linear change of coordinates, (4.5) is transformed into the system

$$\dot{x}_1 = A_{11}x_1 + B_1u + A_{12}x_2 + h_1(u, x_1, x_2), \tag{4.6}$$

$$\dot{x}_2 = A_{22}x_2 + h_2(u, x_1, x_2) \tag{4.7}$$

in which $\dim (x_1) = d$, $\dim (x_2) = n - d$ and

$$\text{rank } [B_1, A_{11}B_1, \cdots, A_{11}^{d-1}B_1] = d. \tag{4.8}$$

Next we introduce the d-parameter family of controls

$$u(\xi, t) = \xi_1u_1(t) + \xi_2u_2(t) + \cdots + \xi_du_d(t). \tag{4.9}$$

For $|\xi|$ restricted small, one has available the corresponding unique solution of

(4.6)–(4.7) defined on $0 \leqq t \leqq 1$:

$$x_1 = X_1(u(\xi, \cdot), t), \tag{4.10}$$

$$x_2 = X_2(u(\xi, \cdot), t) \tag{4.11}$$

for which $X_1(u(\xi, \cdot), 0) = 0$, $X_2(u(\xi, \cdot), 0) = 0$. By differentiation of (4.6)–(4.7) with respect to the parameter ξ, we see that

$$\begin{aligned} \frac{\partial}{\partial t}\left(\frac{\partial X_1}{\partial \xi}\right) &= A_{11}\left(\frac{\partial X_1}{\partial \xi}\right) + A_{12}\left(\frac{\partial X_2}{\partial \xi}\right) \\ &\quad + B_1[u_1(t), \cdots, u_d(t)] + \frac{\partial}{\partial \xi} h_1(u(\xi, t), X_1, X_2), \end{aligned} \tag{4.12}$$

$$\frac{\partial}{\partial t}\left(\frac{\partial X_2}{\partial \xi}\right) = A_{22}\left(\frac{\partial X_2}{\partial \xi}\right) + \frac{\partial}{\partial \xi} h_2(u(\xi, t), X_1, X_2). \tag{4.13}$$

In particular, when $\xi = 0$,

$$\frac{d}{dt}\left(\frac{\partial X_1}{\partial \xi}\right)_0 = A_{11}\left(\frac{\partial X}{\partial \xi}\right)_0 + B_1[u_1(t), u_2(t), \cdots, u_d(t)], \tag{4.14}$$

$$\left(\frac{\partial X_2}{\partial \xi}\right)_0 = 0 \qquad \text{for } 0 \leqq t \leqq 1. \tag{4.15}$$

In view of (4.8), the $u_1(t), u_2(t), \cdots, u_d(t)$ can be chosen such that

$$\left.\frac{\partial X_1(u(\xi, \cdot), 1)}{\partial \xi}\right|_{\xi = 0} = I_d. \tag{4.16}$$

By the implicit function theorem, it follows that (4.10) at $t = 1$,

$$x_1 = X_1(u(\xi, \cdot), 1), \tag{4.17}$$

can be solved for $\xi = \xi(x_1)$ for all x_1 near the origin and $\xi(0) = 0$. Substitution of $\xi(x_1)$ at $t = 1$ into (4.11) shows

$$x_2 = X_2(u(\xi(x_1), \cdot), 1), \tag{4.18}$$

which is the required d-dimensional submanifold of attained states.

Remark 4.1. Generally the manifold discussed in Theorem 4.2 is a proper subset of C. Furthermore, the manifold is not always unique since the control selected to satisfy (4.16) is not unique. We illustrate these remarks by computing the manifolds for Examples 3.1 and 3.2.

In Example 3.1, (4.17) and (4.11) give

$$x_1 = \xi \int_0^1 u(\tau)\, d\tau, \tag{4.19}$$

$$x_2 = X_2(u(\xi, \cdot), 1) = \xi^3 \int_0^1 \left[\int_0^\sigma u(\tau)\, d\tau\right] u^2(\sigma)\, d\sigma \tag{4.20}$$

which can be solved for the manifold (4.18):

$$x_2 = \frac{x_1^3}{\left(\int_0^1 u(\tau)\,d\tau\right)^3}\int_0^1 \left[\int_0^\sigma u(\tau)\,d\tau\right] u^2(\sigma)\,d\sigma. \tag{4.21}$$

By varying the choice of control $u(\cdot)$, this family can be shown to sweep out $C = R^2$.

For Example 3.2, (4.17) and (4.11) give

$$x_1 = \xi\int_0^1 u(\tau)\,d\tau, \tag{4.22}$$

$$\begin{aligned} x_2 &= \xi^2\int_0^1 \left(\int_0^\sigma u(\tau)\,d\tau\right) u(\sigma)\,d\sigma \\ &= \frac{1}{2}\xi^2\left(\int_0^1 u(\tau)\,d\tau\right)^2 \end{aligned} \tag{4.23}$$

which determine the manifold (4.18),

$$x_2 = \tfrac{1}{2}x_1^2, \tag{4.24}$$

which is exactly C.

5. Extensions. The results of Theorems 1.1, 1.2 and 4.1 can easily be extended to system (1.1) in which $A = A(t)$ and $B = B(t)$ are time varying matrices satisfying the generalization of (4.2) (see [4, p. 125]) making the linear part of (1.1) controllable.

REFERENCES

[1] E. J. Davison and E. G. Kunze, *Some sufficient conditions for the global and local controllability of nonlinear time-varying systems*, this Journal, 8 (1970), pp. 489–497.

[2] G. W. Haynes and H. Hermes, *Nonlinear controllability via Lie theory*, this Journal, 8 (1970), pp. 450–460.

[3] H. Hermes, *Controllability and the singular problem*, this Journal, 2 (1965), pp. 241–260.

[4] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[5] D. L. Lukes, *Stabilizability and optimal control*, Funkcial. Ekvac., 11 (1968), pp. 39–50.

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

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

HIGH ORDER NECESSARY CONDITIONS FOR OPTIMALITY*

R. GABASOV AND F. M. KIRILLOVA†

Abstract. This is a survey paper. First, we review various methods of evaluating singular controls. An original proof of Kelley's second order necessary conditions for optimality is given. High order optimality conditions are obtained by the method of Kopp and Moyer. The latter method is generalized to multidimensional singular controls (Goh's result). A method of transformations in state space is presented (Kelley, Gurman). A method for investigating singular controls with the aid of a bundle of variations is described. A survey is made of the results which were obtained with the aid of a bundle of variations and matrix impulses in problems with closed control regions. The problem of joining extremals is singled out. The junction of singular and nonsingular extremals is considered. A survey of necessary conditions for optimality of nonsingular controls is presented. The cases of open and closed control regions are considered separately.

INTRODUCTION

During the fifteen years since the appearance of the first works on the optimization of dynamic control systems, the theory of optimal processes has attained a high stage of development. It has reached its most complete form in the area which deals with necessary conditions for optimality. The basic efforts of many Soviet and non-Soviet scientists were directed toward generalizing and developing Pontryagin's maximum principle, which was discovered in 1956 [1], [2]. This result, which is a substantial generalization of the classical necessary conditions of the calculus of variations [3], gave a powerful push for developing optimization methods. The huge and still increasing role of these methods in modern technology is universally recognized.

The first systems for which the theory of optimal processes began to be constructed were ordinary dynamic systems (in other terms, lumped parameter systems). The behavior of each system is described by an ordinary differential equation

$$dx/dt = f(x(t), u(t), t), \tag{0.1}$$

where $x = \{x_1, \cdots, x_n\}$ is an n-dimensional state vector (of phase coordinates), $u = \{u_1, \cdots, u_r\}$ is an r-dimensional control vector, and t is the time.

Given a particular class of admissible controls (usually, this is the set of all piecewise-continuous or measurable functions which take on their values in some set U), and having chosen an optimality criterion, one can formulate various optimization problems. In these problems, various constraints on the state and on the controls may additionally be taken into account. The spectrum of such problems and of methods for solving them (more precisely, of proofs of the maximum principle for them) is very rich [4]–[12].

The problem which lies at the foundation of a possible direction of developing necessary conditions for optimality consists of investigating high order optimality conditions. By these, one understands those properties which distinguish the

* Received in Russian by the editors on April 10, 1970, and in final revised form on January 26, 1971. This translation into English has been prepared by K. Makowski.

Translated and printed for this Journal under a grant-in-aid from the National Science Foundation.

† Mathematics Institute of the Academy of Sciences of the Byelorussian SSR, Minsk, USSR.

optimal controls (generally speaking) from the remaining controls which satisfy the maximum principle (or its consequences, the first order optimality conditions). A more precise definition of this new concept will be given in the textual part of the survey.

In this paper, a survey of some investigations in this direction is given. Since all the results of this type could not have arisen without an investigation of the simpler first order optimality conditions, we have touched upon works concentrated around the maximum principle. In the sequel, we shall only talk about those works in which conditions for optimal controls which supplement the maximum principle and its consequences are proved. The theme of the survey is a comparatively new one; therefore, the aim of the survey will also differ significantly from the aims of traditional surveys. The authors have not attempted to summarize the results of some stage of development, but rather have tried to formulate, or to direct attention toward, new problems in optimal control theory. High order necessary conditions for optimality, as such, have not been considered in the literature on optimal processes. The classical second order conditions of the calculus of variations are well known. Similar conditions in the theory of optimal processes were related only to singular controls. In this survey, optimality conditions for singular controls form a part of the high order conditions. The remainder is devoted to necessary conditions for optimality of controls which satisfy the maximum principle without any singularities.

Of the surveys on singular controls, [13] and [14] were available to us. In [14], a great deal of attention was paid only to the one-dimensional case. Also, a large amount of material was not included in [13].

In this survey, as a rule, sliding regimes are not singled out as independent subjects for investigation, but they are considered to be a part of the theory of singular controls. The transition from the former to the latter has been investigated in the literature in detail in [15] and [16].

If, in a problem of minimizing the functional

$$J(u) = \phi(x(t_1)), \qquad t_1 = \text{const.}, \quad t_1 > t_0, \tag{0.2}$$

subject to (0.1), an optimal control does not exist in the class of measurable admissible controls (but is realized as a sliding regime), then this problem is replaced by an auxiliary one in which, instead of (0.1), one considers the following system:

$$\frac{dx}{dt} = \sum_{i=1}^{n+1} \alpha_i(t) f(x(t), u_i(t), t), \tag{0.3}$$

$$\alpha_i(t) \geqq 0, \quad \sum_{i=1}^{n+1} \alpha_i(t) = 1, \quad u_i(t) \in U. \tag{0.4}$$

The new problem with an $(n + 1)(r + 1)$-dimensional control $\{\alpha, u\}$ has a solution under very general conditions [17]. However, it is impossible to find this solution with the aid of the maximum principle because the solution contains the singular components α_i on which the function H does not depend along an optimal control. The transition from the problem (0.3), (0.4) and (0.2) to the original one has also been well developed in the literature.

The basic assertions in this paper are accompanied by examples which have an illustrative character. In the literature on singular controls, very interesting problems, which have a practical side to them, were computed. Sources are indicated in the corresponding places of this survey, but the authors considered it inappropriate to present these and similar problems as examples, since this would have significantly influenced the size of this survey, and would have clouded the basic idea of the mathematical result with additional difficulties which are not, generally speaking, peculiar to the result itself. Finally, the majority of the known examples of this kind are related to problems of rocket dynamics, and, therefore, are of interest only to a certain circle of experts. From our point of view, already solved optimization problems deserve a special survey.

In writing this survey, the authors, following a widespread practice in such cases, have considered it appropriate to frequently omit mentioning those analytic properties of the functions (of smoothness type) under which some assertion or other holds. In the majority of cases, the class of functions being used is obvious, or is generally accepted in the optimal control literature. One can find a precise description of these properties in the corresponding sources on whose basis the survey has been made. For the sake of simplicity in the presentation, all of the functions being used will be assumed to be sufficiently smooth so that all of the operations performed on them are valid. Such a "simplification" has not been made with regard to other objects which are encountered in the survey. For example, conditions regarding whether sets are closed or convex are noted everywhere since, from our viewpoint, such assumptions are essential from the very nature of the problem rather than being related to the level of the mathematical techniques being used.

We present few words on the notation. Vectors are everywhere to be understood to have been written as column-vectors. To denote row-vectors, the symbol $'$ (prime) is used. This allows us to write the scalar product of vectors a and b as $a'b$. If x is a scalar, the symbol $\partial S/\partial x$ will in this survey denote the ordinary derivative of a function S with respect to x. If x is a vector, then $\partial S/\partial x = \operatorname{grad} S$ will denote the gradient of the function S, written as a column vector. The symbol $\partial f/\partial x$, where f and x are n-vectors, will denote the matrix consisting of the elements $\alpha_{ij} = \partial f_i/\partial x_j$, $i = 1, \cdots, n$, $j = 1, \cdots, n$.

SINGULAR CONTROLS

1. Formulation of the problem. Definitions of singular controls.

1.1. Let us consider a typical optimization problem. We are given the system

$$\dot{x} = f(x, u, t), \quad x(t_0) = x_0, \quad t \in T = [t_0, t_1], \tag{1}$$

$$x = \{x_1, \cdots, x_n\}, \qquad u = \{u_1, \cdots, u_r\}.$$

Using r-dimensional, piecewise-continuous functions $u(t)$ taking on their values in a bounded set U (admissible controls),

$$u(t) \in U, \qquad t \in T, \tag{2}$$

we want to minimize the functional

$$J(u) = \phi(x(t_1)), \tag{3}$$

defined on the trajectories of system (1).

Let

$$H(x, \psi, u, t) = \psi' f(x, u, t) \tag{4}$$

be the Hamiltonian, and let

$$\begin{aligned} &\dot{x} = \partial H/\partial \psi, \qquad x(t_0) = x_0, \\ &\dot{\psi} = -\frac{\partial H}{\partial x}, \qquad \psi(t_1) = -\frac{\partial \phi(x(t_1))}{\partial x} \end{aligned} \tag{5}$$

be the canonical equations for the phase vector x and the impulse $\psi = \{\psi_1, \cdots, \psi_n\}$ (the adjoint variable vector).

It is well known (see the Introduction) that the optimal controls for problem (1)–(3) are found among the admissible controls $u(t)$, $t \in T$, which satisfy the maximum principle

$$H(x(t), \psi(t), v, t) - H(x(t), \psi(t), u(t), t) \equiv \Delta_v H(x(t), \psi(t), u(t), t) \leqq 0 \tag{6}$$

for all $v \in U$,

where $x(t), \psi(t)$ is a solution of system (5) which corresponds to a control $u(t)$.

If U is an open set, it follows from (6) that

$$\frac{\partial H}{\partial u} = 0, \qquad \eta' \frac{\partial^2 H}{\partial u^2} \eta \leqq 0. \tag{7}$$

DEFINITION 1. A control $u(t)$ is said to be *singular* (*in the sense of the maximum principle* (6)) if there exists a nontrivial set $\omega(t) \subset U$ such that

$$\Delta_v H(x(t), \psi(t), u(t), t) \equiv 0 \quad \text{for all } v \in \omega(t) \quad \text{and} \quad t \in T. \tag{8}$$

DEFINITION 2. A control $u(t)$ is said to be *singular* (*in the classical sense*) if the following identities hold:

$$\frac{\partial H(x(t), \psi(t), u(t), t)}{\partial u} \equiv 0, \qquad \det \left| \frac{\partial^2 H(x(t), \psi(t), u(t), t)}{\partial u^2} \right| \equiv 0 \quad \text{for all } t \in T. \tag{9}$$

1.2. If U is open and $\omega(t) = U$, then it is clear that every control which is singular in the sense of the maximum principle is also singular in the classical sense. The first definition is used when an optimal control lies on the boundary of U and the Hamiltonian is not stationary at $u(t)$. If an optimal control is in the interior of U, then either definition may be used. But, since the second definition is broader than the first one (it is satisfied by more controls, including ones which may not be singular in the case of the maximum principle), it turns out to be useful in obtaining higher order optimality conditions.

Definition 1 is closest to the definitions of singular controls given in [11], [18] and [19]. The second definition is more often used in the literature. For systems of the form

$$\dot{x} = g(x, t) + \sum_{\nu=1}^{r} h_\nu(x, t) u_\nu,$$

which were studied by most previous authors, these definitions coincide. For singular controls in such systems, there always exists a subset ω of $\{1, \cdots, r\}$ such that

$$\psi'(t)h_\mu(x(t), t) \equiv 0 \quad \text{for all } \mu \in \omega \quad \text{and} \quad t \in T.$$

Remark. Here and in the sequel, if no special statement to the contrary is made, controls will be assumed to be singular on the entire interval T. Obviously, going over to the particular case where $u(t)$ is singular on an interval $\sigma \subset T$ causes no difficulties.

2. Methods of evaluating singular controls.

2.1. Some existence conditions and some methods of evaluating singular controls follow directly from Definitions 1 and 2 and from constraint (2). For a singular control, identities (8) and (9) must hold. Thus, if we successively differentiate these identities with respect to time, taking into account (5), we are led, generally speaking, to a system of equations (which are algebraic and differential) for the components of a singular control. The problems of existence and of how to evaluate a singular control therefore reduce to the question of whether the system thus obtained can be solved subject to constraint (2). In the general case, such a procedure (which is very cumbersome) leads to a singular control which depends on the variables $x, \psi,$ and t: $u = u(x, \psi, t)$.

Example 1.

$$\dot{x}_1 = u, \quad \dot{x}_2 = -x_1^2 + u^4, \quad x_1(0) = x_2(0) = 0, \quad T = [0, 1],$$
$$|u| < 1, \qquad \phi(x) = x_2,$$
$$H(x, \psi, u) = \psi_1 u + x_1^2 - u^4, \quad \dot{\psi}_1 = -2x_1, \quad \psi_1(1) = 0.$$

According to Definition 2, we obtain

$$\partial H/\partial u = \psi_1 - 4u^3 = 0, \qquad \partial^2 H/\partial u^2 = -12u^2 = 0.$$

Thus, the singular control has the form $u = 0$, and it is admissible.

2.2. Methods of evaluating singular controls have been investigated most completely for systems of the form

$$\dot{x} = f_0(x) + uf_1(x), \qquad u \in U, \tag{10}$$

with a single control [13], [20] and [21]. Here, a control $u(t)$ is singular if and only if the Hamiltonian is stationary:

$$\partial H/\partial u = \psi'(t)f_1(x(t)) \equiv 0, \qquad t \in T, \tag{11}$$

where $\psi(t)$ is a solution of the equation

$$\dot{\psi} = -\frac{\partial f_0'}{\partial x}\psi - u\frac{\partial f_1'}{\partial x}\psi. \tag{12}$$

Following our general scheme (see § 2.1), we shall successively evaluate the total derivatives $(d^m/dt^m)(\partial H/\partial u), m = 1, 2, \cdots$, along the trajectories of system (10), (12). The general structure of these derivatives is described in [20]. The first value

of the index m for which the control u appears in $(d^m/dt^m)(\partial H/\partial u)$ must be even [14], and for this $m = 2k$, the control must enter this expression linearly. Thus, we have

$$\frac{d^m}{dt^m}\frac{\partial H}{\partial u} = 0, \qquad m = 0, 1, \cdots, 2k-1,$$

(13)

$$\frac{d^{2k}}{dt^{2k}}\frac{\partial H}{\partial u} \equiv a(x,\psi) + b(x,\psi)u = 0,$$

$$b(x,\psi) \neq 0, \quad k = 1, 2, \cdots.$$

Hence,

$$u(x,\psi) = -a(x,\psi)/b(x,\psi). \tag{14}$$

This singular control is admissible for those x and ψ for which the right-hand side of (14) belongs to U.

Example 2.

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_3 + u, \quad \dot{x}_3 = x_4, \quad \dot{x}_4 = ux_1^2, \quad J(u) = \phi(x_1, x_2, x_3, x_4).$$

Here,

$$H = \psi_1 x_2 + \psi_2(x_3 + u) + \psi_3 x_4 + \psi_4 u x_1^2,$$

$$\dot{\psi}_1 = -2\psi_4 x_1 u, \quad \dot{\psi}_2 = -\psi_1, \quad \dot{\psi}_3 = -\psi_2, \quad \dot{\psi}_4 = -\psi_3.$$

For a singular control, the following must hold:

$$\frac{\partial H}{\partial u} = \psi_2 + \psi_4 x_1^2 = 0,$$

$$\frac{d}{dt}\frac{\partial H}{du} = -\psi_1 - \psi_3 x_1^2 + 2\psi_4 x_1 x_2 = 0,$$

$$\frac{d^2}{dt^2}\frac{\partial H}{\partial u} = \psi_2 x_1^2 - 4\psi_3 x_1 x_2 + 2\psi_4 x_2^2 + 2\psi_4 x_1 x_3 + 4\psi_4 x_1 u = 0.$$

The singular control is determined by the last of these equations so long as $\psi_4 x_1 \neq 0$:

$$u = \frac{4\psi_3 x_1 x_2 + \psi_4 x_1^4 - 2\psi_4 x_2^2 - 2\psi_4 x_1 x_3}{4\psi_4 x_1}.$$

2.3. In some cases, the form of the singular control (14) may be simplified. Let us define the operation

$$[f, g] = \frac{\partial f}{\partial x} g - \frac{\partial g}{\partial x} f,$$

where $f(x)$ and $g(x)$ are n-dimensional vector-valued functions. Then, equations (13)

take the form

$$\frac{\partial H}{\partial u} = \psi' f_1 = 0,$$

$$\frac{d}{dt}\frac{\partial H}{\partial u} = -\psi'[f_0,f_1] = 0, \tag{15}$$

$$\cdots$$

$$\frac{d^m}{dt^m}\frac{\partial H}{\partial u} = (-1)^m \psi'[\underbrace{f_0,[f_0,[\cdots,[f_0}_{m},f_1]]]\cdots] = 0, \qquad m < 2k.$$

If $m < n-1$, then $m+1$ components of ψ can be eliminated from (14) with the aid of (15), and the singular control will not depend on these components. If $m \geqq n-1$, then in system (15), which is linear with respect to ψ, by virtue of the nontriviality of the vector ψ (according to the maximum principle), at least one nth order minor $S(x)$ of the determinant

$$|f_1 \vdots -[f_0,f_1] \vdots \cdots \vdots (-1)^m[\underbrace{f_0,[f_0,[\cdots,[f_0}_{m},f_1]]]\cdots]|$$

vanishes [22]:

$$S(x) = 0. \tag{16}$$

The set of points x which satisfy (16) generates the singular control surface S (see [21]) which contains the trajectories of system (10) which correspond to singular controls. In this case, the singular controls are determined by the condition that the trajectories lie on S:

$$\frac{dS(x)}{dt} = \frac{\partial S'(x)}{\partial x}(f_0(x) + uf_1(x)) = 0. \tag{17}$$

Hence, if $(\partial S'(x)/\partial x)f_1(x) \neq 0$, we obtain the following expression for a singular control:

$$u(x) = -\frac{(\partial S'(x)/\partial x)f_0(x)}{(\partial S'(x)/\partial x)f_1(x)}. \tag{18}$$

Since the minimum number of equations in (15) is two, the equation for S can always be obtained from system (10) with $n = 1, 2$ (see [21]).

Example 3.

$$\dot{x}_1 = u, \quad \dot{x}_2 = -x_1^2, \quad x_1(0) = x_2(0) = 0, \quad T = [0,1], \quad \phi(x) = x_2, \quad |u| < 1,$$

$$H(x,\psi,u) = \psi_1 u + x_1^2, \quad \dot{\psi}_1 = -2x_1, \quad \psi_1(1) = 0,$$

$$\frac{\partial H}{du} = \psi_1 = 0, \quad \frac{d}{dt}\frac{\partial H}{\partial u} = -2x_1 = 0.$$

We thus conclude that the singular control surface is characterized by the equation $x_1 = 0$, and, by virtue of (18), the singular control has the form $u = 0$.

2.4. The derivatives $(d^m/dt^m)(\partial H/\partial u)$, $m = 1, 2, \cdots$, which we need in order to evaluate a singular control, can be expressed in a suitable form with the aid

of Poisson brackets. Let

$$v = v(q_1, \cdots, q_n, p_1, \cdots, p_n), \qquad w = w(q_1, \cdots, q_n, p_1, \cdots, p_n).$$

The expression

$$\{v, w\} = \sum_{i=1}^{n} \left(\frac{\partial v}{\partial q_i} \frac{\partial w}{\partial p_i} - \frac{\partial v}{\partial p_i} \frac{\partial w}{\partial q_i} \right)$$

is said to be the Poisson bracket of the functions v and w (see [23]). If now $S = S(x, \psi)$, then the derivative with respect to time of this function along trajectories of system (10), (12) can be written in the form

$$dS/dt = \{S, H\}, \tag{19}$$

where H is the Hamiltonian function (4).

Let

$$H(x, \psi, u) = H_0(x, \psi) + uH_1(x, \psi), \tag{20}$$

where $H_0 = \psi' f_0$ and $H_1 = \psi' f_1$. Taking into account (19), we can express the derivatives (13) in the following form:

$$\begin{aligned}
&\frac{\partial H}{\partial u} = H_1 = 0, \\
&\frac{d}{dt}\frac{\partial H}{\partial u} = -\{H_0, H_1\} = 0, \qquad \{H_1, H_2\} \equiv 0, \\
&\frac{d^2}{dt^2}\frac{\partial H}{\partial u} = \{H_0, \{H_0, H_1\}\} = 0, \qquad \{H_1, \{H_0, H_1\}\} \equiv 0, \\
&\cdots \\
&\frac{d^m}{dt^m}\frac{\partial H}{\partial u} = (-1)^m \{\underbrace{H_0, \{H_0, \cdots, \{H_0}_{m}, H_1\}\} \cdots\} = 0,
\end{aligned} \tag{21}$$

$$\{H_1, \{\underbrace{H_0, \cdots, \{H_0}_{m-1}, H_1\}\} \cdots\} \equiv 0, \quad m < 2k, \tag{22}$$

$$\frac{d^{2k}}{dt^{2k}}\frac{\partial H}{\partial u} = \{\underbrace{H_0, \{H_0, \cdots, \{H_0}_{2k}, H_1\}\} \cdots\} + \{H_1, \{\underbrace{H_0, \cdots, \{H_0}_{2k-1}, H_1\}\} \cdots\} u = 0.$$

A singular control is determined by the last equation. The meaning of the coefficient of u in this equation

$$\frac{\partial}{\partial u}\left(\frac{d^{2k}}{dt^{2k}}\frac{\partial H}{\partial u}\right) = \{H_1, \{H_0, \{\underbrace{H_0, \cdots, \{H_0}_{2k-1}, H_1\}\}\} \cdots\}, \qquad k = 1, 2, \cdots, \tag{23}$$

will be clarified in the sequel (see §§ 3 and 4).

Thus, the derivatives $(d^m/dt^m)(\partial H/\partial u)$ (for finding a singular control) can be successively evaluated with the aid of the operation of taking the Poisson brackets of the two functions H_0 and H_1 which enter in the Hamiltonian (20).

Remark. Using the representation (21) and the Jacobi identity for the Poisson brackets [23], it is easy to show that the first value of m for which a control can appear in a derivative $(d^m/dt^m)(\partial H/\partial u)$ must be even. This fact was proved in [14] by other considerations.

2.5. In the monograph [24], it was proposed that one should use Poisson's theorem to evaluate singular controls. For a singular control of system (10), the condition

$$H_1(x, \psi) = 0 \tag{24}$$

holds. Since the Hamiltonian (20) is constant along a singular extremal,

$$H_0(x, \psi) = C. \tag{25}$$

In [24], equations (24) and (25) were interpreted as two first integrals of the canonical system (10), (12). By Poisson's theorem [23], the Poisson bracket $\{H_0, H_1\}$, which is made up of the two first integrals, is also a first integral of the canonical system. In other words, the following condition holds for a singular control:

$$\{H_0, H_1\} = 0.$$

This is equivalent to the second of equations (21). In such a way, it is also possible to obtain all of the remaining equations (21). Choosing a linearly independent collection from the first integrals thus obtained, the author of [24] arrived at a system of equations which is linear in ψ and analogous to (15). Taking into account the fact that the matrix of coefficients of this system is singular, in [24], equation (16) for the singular control surface was obtained. The singular control (18) is determined by condition (17). If, according to the procedure described in § 2.3, only equations (21), which do not contain u, are used to evaluate the singular control surface, then in [24] the remaining equations (22) were also used to obtain S, but the coefficients of u in these equations are set equal to zero. This operation means that only those singular controls on which these coefficients really vanish are being sought. In other words, singular controls of the form (18) which admit the singular control surface (16) are being sought.

Following the described procedure, we shall find S and a singular control in Example 2. In this case, it is sufficient to take four first integrals. They are of the form

$$H_1 = \psi_2 + \psi_4 x_1^2 = 0, \qquad \{H_0, H_1\} = \psi_1 + \psi_3 x_1^2 - 2\psi_4 x_1 x_2 = 0,$$
$$\{H_0, \{H_0, H_1\}\} = \psi_2 x_1^2 - 4\psi_3 x_1 x_2 + 2\psi_4(x_2^2 + x_1 x_3) = 0,$$
$$\{H_1, \{H_0, H_1\}\} = 4\psi_4 x_1 = 0.$$

Setting the determinant of this system (which is linear in ψ) equal to zero, we obtain the equation for S: $x_1 = 0, x_2 = 0$. According to (18), the singular control has the form $u = -x_3$.

In [24], the procedure of this section is applied to a problem of flying an aircraft to a given altitude.

2.6. The possibility that a singular control may appear in the system

$$\dot{x} = Ax + bu, \qquad |u| \leqq 1, \tag{26}$$

with the functional

$$J(u) = \int_{t_0}^{t_1} x'Qx\,dt \tag{27}$$

has been investigated in [25]. In this case, S is the hyperplane

$$C'x = 0, \tag{28}$$

and the singular control is linear in x: $u = \gamma'x$. The vectors C and γ may be expressed in terms of the parameters of the system.

A singular control in systems with a variable structure:

$$\begin{aligned} &\dot{x}_i = x_{i+1}, \qquad i = 1, \cdots, n-1, \\ &\dot{x}_n = \alpha'x + uk'x, \\ &J(u) = \int_0^{\infty} x'Px\,dt, \end{aligned} \tag{29}$$

has been investigated in [26] and [27]. It was shown in these works that the singular control surface has the form (28), and the singular control may be represented as follows: $u = m'x/b'x$, $b'x \neq 0$.

The work [28] is devoted to finding a multidimensional singular control in the problem

$$\dot{x} = Ax + Bu, \qquad u = \{u_1, \cdots, u_r\},$$

$$J(u) = g(x(t_1), t_1) + \int_{t_0}^{t_1} (u'Ru + 2u'Qx + x'Px)\,dt.$$

Remark 1. In order to find singular multidimensional controls according to the general procedure described in § 2.1, one must take into account the necessary conditions for optimality in the form of equations which hold in this case (see § 5).

Remark 2. In problems with a free time t_1, to evaluate a singular control one can use the additional relation $H(x, \psi, u) \equiv 0$, which holds for an optimal control.

Remark 3. The above described methods do not allow one to find all of the singular controls which can be encountered in an optimization problem.

Example 4.

$$\dot{x}_1 = u, \quad \dot{x}_2 = ux_1, \quad x_1(0) = x_2(0) = 0, \quad |u| \leqq 1,$$

$$T = [0, 1], \qquad \phi(x) = x_2.$$

Here,

$$H = \psi_1 u + ux_1, \quad \dot{\psi}_1 = -u, \quad \psi_1(1) = 0,$$

$$\frac{\partial H}{\partial u} = \psi_1 + x_1 = 0, \qquad \frac{d}{dt}\frac{\partial H}{\partial u} = u - u \equiv 0.$$

All of the higher derivatives vanish. Therefore, using the methods described in this paragraph, it is impossible to find either the singular control ($u = 0$), or the

singular control surface ($x_1 = 0$). Since there is a finite number of optimal controls in this example, the assumption in [22, p. 437] is not satisfied.

Remark 4. Singular controls are also encountered in the problems of minimizing the functional [22]

$$J(u) = \int_{t_0}^{t_1} [f_{00}(x(t)) + f_{01}(x(t))|u(t)|]\,dt, \qquad |u(t)| \leqq 1, \tag{30}$$

on the trajectories of system (10). (In particular cases, this reduces to a problem of minimum fuel or mass expenditure, etc.) It is not difficult to see that, in such problems, a singular control is characterized by one of the identities

$$\psi'(t)f_1(x(t)) \pm f_{01}(x(t)) \equiv 0, \qquad t \in T. \tag{31}$$

If we adjoin an $(n + 1)$st component $\psi_0 = \pm 1$ to the impulse n-vector ψ, then the property (31) becomes equivalent to (11). Thus, the singular controls $|u(t)| < 1$ in problem (10), (30) reduce to the controls investigated in § 2.2–§ 2.5, and all of the conclusions of these sections hold for these controls. The only complication that arises is caused by the fact that the order of the system is increased by 1, and therefore it becomes necessary to consider two cases: $\psi_0 = 1$ and $\psi_0 = -1$.

Remark 5. Problems of evaluating singular controls have also been considered in [29]–[33].

3. Kelley's method of obtaining necessary conditions for optimality for a singular control. The one-dimensional control case.

3.1. A general method of deriving necessary conditions for optimality, based on an investigation of the second variation of the functional being minimized, was proposed by Kelley in [34]. The considerations in [34] were carried out for the problem (1)–(3), where u is a scalar function and U is an open set.

Let us describe the basic steps of the method. The formula for an increment of the functional (3) has the form [35]

$$\begin{aligned}
\Delta J(u) = &- \int_{t_0}^{t_1} [H(x, \psi, u + \delta u, t) - H(x, \psi, u, t)]\,dt \\
&- \int_{t_0}^{t_1} \Delta x'(t) \frac{\partial [H(x, \psi, u + \delta u, t) - H(x, \psi, u, t)]}{\partial x}\,dt \\
&- \frac{1}{2}\int_{t_0}^{t_1} \Delta x'(t) \frac{\partial^2 H(x, \psi, u + \delta u, t)}{\partial x^2} \Delta x(t)\,dt \\
&+ \frac{1}{2}\Delta x'(t_1) \frac{\partial^2 \phi(x(t_1))}{\partial x^2} \Delta x(t_1) \\
&- \int_{t_0}^{t_1} o_1(\|\Delta x(t)\|^2)\,dt + o(\|\Delta x(t_1)\|^2).
\end{aligned} \tag{32}$$

Separating out in (32) the principal terms, up to and including the terms of second order with respect to $\delta u(t)$, we obtain

$$\Delta J(u) = \delta J + \tfrac{1}{2}\delta^2 J + o(|\delta u(t)|^2),$$

where

$$\delta J = -\int_{t_0}^{t_1} \frac{\partial H}{\partial u}\,\delta u\,dt,$$

$$\delta^2 J = -\int_{t_0}^{t_1} \left[\frac{\partial^2 H}{\partial u^2}\delta u^2 + 2\frac{\partial^2 H'}{\partial x \partial u}\delta x \delta u + \delta x' \frac{\partial^2 H}{\partial x^2}\delta x\right] dt$$

$$+ \delta x'(t_1)\frac{\partial^2 \phi(x(t_1))}{\partial x^2}\delta x(t_1). \tag{33}$$

Here, $\delta x(t)$ is a solution of the variational equation

$$\delta \dot{x} = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial u}\delta u, \qquad \delta x(t_0) = 0. \tag{34}$$

Expression (33) represents the second variation of the functional (3). For an optimal control, $\delta J = 0$ and the inequality $\delta^2 J \geqq 0$ follows from the condition $\Delta J(u) \geqq 0$. This inequality is used to obtain necessary conditions for optimality of the singular controls.

For a control which is singular in the classical sense, the second variation has the form

$$\delta^2 J = -\int_{t_0}^{t_1} \left(2\frac{\partial^2 H'}{\partial u \partial x}\delta x \delta u + \delta x' \frac{\partial^2 H}{\partial x^2}\delta x\right) dt$$

$$+ \delta x'(t_1)\frac{\partial^2 \phi(x(t_1))}{\partial x^2}\delta x(t_1) \tag{35}$$

and the classical methods of investigation [3] are not applicable here.

3.2. Expression (35) was investigated in [34] for variations in the controls of the form

$$\delta u(t) = \begin{cases} a, & \theta \leqq t < \theta + \varepsilon, \\ -a, & \theta + \varepsilon \leqq t < \theta + 2\varepsilon, \\ 0, & \text{for remaining } t \in T, \end{cases} \tag{36}$$

where $\theta \in T$, $\varepsilon > 0$, and a is an arbitrary number.

For the variations (36), a solution of (34) has the property that

$$\delta x(t) = o(\varepsilon^2), \qquad \theta + 2\varepsilon \leqq t \leqq t_1. \tag{37}$$

This property substantially simplifies the structure of the second variation. It is clear that, by virtue of (37), the expression

$$-\int_{\theta+2\varepsilon}^{t_1} \delta x' \frac{\partial^2 H}{\partial x^2}\delta x\,dt + \delta x'(t_1)\frac{\partial^2 \phi}{\partial x^2}\delta x(t_1)$$

in (35) is of order ε^4. The principal term in the expansion of $\delta^2 J$ in powers of ε is of order ε^3, and is obtained by considering the integral

$$-\int_{\theta}^{\theta+2\varepsilon} \left(2\frac{\partial^2 H'}{\partial u \partial x}\delta x\,\delta u + \delta x' \frac{\partial^2 H}{\partial x^2}\delta x\right) dt.$$

It is not difficult to find that

$$\delta^2 J = \tfrac{2}{3}P(\theta)a^2\varepsilon^3 + o(\varepsilon^3),$$

where

$$P(t) = -\frac{\partial f'}{\partial u}\frac{\partial^2 H}{\partial x^2}\frac{\partial f}{\partial u} + 2\frac{\partial^2 H}{\partial u \partial x}\frac{\partial f}{\partial x}\frac{\partial f}{\partial u} - \frac{\partial^2 H}{\partial u \partial x}\frac{d}{dt}\frac{\partial f}{\partial u} + \frac{d}{dt}\frac{\partial^2 H'}{\partial u \partial x}\frac{\partial f}{\partial u}. \tag{38}$$

Therefore, for a singular control to be optimal, it is necessary that

$$P(t) \geqq 0, \qquad t \in T. \tag{39}$$

This condition, obtained by Kelley [34], can be presented (Bryson) in compact form as follows:

$$\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) \geqq 0. \tag{40}$$

On the other hand, it follows from § 2.4 that condition (39) can also be written in terms of Poisson brackets as follows:

$$P(t) = \left\{\frac{d}{dt}\frac{\partial H}{\partial u}, \frac{\partial H}{\partial u}\right\} \geqq 0.$$

It is interesting to note that, for systems which are linear in the control, the left-hand side of (40) is the coefficient in (23) for $k = 1$ of u in equation (22), from which a singular control can be found. Thus, one can simultaneously evaluate a singular control according to the procedure described in § 2.4 and verify whether this control is optimal with the aid of condition (40): If the coefficient of u in $(d^2/dt^2)(\partial H/\partial u)$ does not vanish, it must be positive. But if, while evaluating a singular control, this coefficient turns out to be zero, then, for the singular controls which are found from subsequent investigations, condition (40) is not effective. From this it follows that, for the singular controls evaluated by the procedure of § 2.5, in systems of order $n \geqq 4$, the necessary condition (40) is always satisfied, with equality holding. We shall illustrate the just-obtained condition by applying it to Example 3. We have

$$\frac{d^2}{dt^2}\frac{\partial H}{\partial u} = -2u, \qquad \frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) = -2 < 0,$$

which contradicts (40). Thus, the singular control $u = 0$ is not optimal.

With the aid of the necessary condition (40), the character of a singular control in various problems has been investigated, such as the Goddard problem [14] and [34], the problem of controlling the middle part of a trajectory [36], and a servomechanism problem [21].

4. The method of transformations in the control space. The one-dimensional control case. To investigate the second variation, in [37] a procedure different than that of [34] was used. This procedure is, in a well-known sense, a generalization of the latter. As was noted in § 3.1 for a singular control, the second variation

has a form to which the classical conditions of Legendre and Clebsch cannot be applied. Therefore, it was proposed in [37] to reduce expression (35), with the aid of transformations on the control variations, to a form which allows one to effectively apply the conditions of Legendre and Clebsch. If, after a single application of such a transformation, which leads to Kelley's condition (39), an ineffective result is obtained (condition (39) holds as an equality), then the transformation is repeated. In this way, we obtain a sequence of the inequality-type necessary conditions for optimality of the singular controls. Each of these conditions is applicable when all of the preceding ones hold as equalities.

Let us briefly describe this procedure. The second variation for a singular control has the form (35), where δu and δx satisfy (34). Let us pass to new variations δv and δy according to the formulas

$$\delta v(t) = \int_{t_0}^{t} \delta u(\tau)\, d\tau, \qquad \delta y(t) = \delta x(t) - \frac{\partial f}{\partial u}\delta v(t). \tag{41}$$

It is not difficult to see that δv and δy satisfy the equation

$$\delta \dot{y} = \frac{\partial f}{\partial x}\delta y + \left(\frac{\partial f}{\partial x}\frac{\partial f}{\partial y} - \frac{d}{dt}\frac{\partial f}{\partial u}\right)\delta v, \qquad \delta y(t_0) = 0. \tag{42}$$

In addition, we set

$$\delta v(t_1) = 0. \tag{43}$$

Then, $\delta^2 J$ has the following form in the new variations (in the transformation to this form $\delta \dot{v}(t)$ is eliminated by integrating by parts):

$$\delta^2 J = \int_{t_0}^{t_1} \left[P(t)\delta v^2 + 2\left(-\frac{\partial^2 H}{\partial x^2}\frac{\partial f}{\partial u} + \frac{d}{dt}\frac{\partial^2 H}{\partial u \partial x} + \frac{\partial f'}{\partial x}\frac{\partial^2 H}{\partial u \partial x}\right)' \delta y\, \delta v - \delta y' \frac{\partial^2 H}{\partial x^2}\delta y \right] dt + \delta y'(t_1)\frac{\partial^2 \phi}{\partial x^2}\delta y(t_1), \tag{44}$$

where $P(t)$ is determined by (38). The condition of Legendre and Clebsch can be applied to the second variation in the form (44), which leads to inequality (39) which holds for an optimal singular control (this condition can be obtained by expanding $\delta^2 J$ in powers of ε, if we choose the variation δv in a "needle" form). The original control variation $\delta u(t)$, by virtue of (43), must satisfy the condition

$$\int_{t_0}^{t_1} \delta u(t)\, dt = 0. \tag{45}$$

This condition is essential for control variations with the aid of which the necessary condition (39) can be obtained. The variation (36), used by Kelley, obviously satisfies condition (45).

Now let condition (39) be ineffective for a singular control, i.e., suppose that $P(t) \equiv 0$. Then we make a change of variations in (44) similar to (41):

$$\delta w = \int_{t_0}^{t} \delta v(\tau)\, dt, \qquad \delta z = \delta y - \left(\frac{\partial f}{\partial x}\frac{\partial f}{\partial u} - \frac{d}{dt}\frac{\partial f}{\partial u}\right)\delta w, \tag{46}$$

where δw and δz are new variations. The equation for these variations follows in an obvious way from (42). From the second variation (44), transformed in accordance with the change (46) ($\delta w(t_1) = 0$), we obtain a new necessary condition. If this condition is also ineffective, the process is continued. The explicit form of the necessary conditions obtained successively by the described procedure is given in [14] and [37]. In a compact form, they are (Robbins)

$$(47) \qquad (-1)^k \frac{\partial}{\partial u}\left(\frac{d^{2k}}{dt^{2k}} \frac{\partial H}{\partial u}\right) \leqq 0, \qquad k = 1, 2, \cdots .$$

The representation (47) has been proved [14] for systems which are linear in the control. For $k = 1$, we obtain Kelley's necessary condition (40). If it is ineffective, we apply the next condition in (47) with $k = 2$, etc.

Example 5.

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad \dot{x}_3 = -x_1^2, \quad x_1(0) = x_2(0) = x_3(0) = 0, \quad |u| < 1,$$

$$T = [0, 1], \quad \phi(x) = x_3, \quad H(x, \psi, u) = \psi_1 x_2 + \psi_2 u + x_1^2,$$

$$\dot{\psi}_1 = -2x_1, \quad \psi_1(1) = 0, \quad \dot{\psi}_2 = -\psi_1, \quad \psi_2(1) = 0.$$

For a singular control,

$$\frac{\partial H}{\partial u} = \psi_2 = 0, \quad \frac{d}{dt}\frac{\partial H}{\partial u} = -\psi_1 = 0, \quad \frac{d^2}{dt^2}\frac{\partial H}{\partial u} = 2x_1 = 0,$$

$$\frac{d^3}{dt^3}\frac{\partial H}{\partial u} = 2x_2 = 0, \qquad \frac{d^4}{dt^4}\frac{\partial H}{\partial u} = 2u = 0.$$

Thus, $u = 0$ is a singular control. For this control,

$$\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) = 0, \qquad \frac{\partial}{\partial u}\left(\frac{d^4}{dt^4}\frac{\partial H}{\partial u}\right) = 2 > 0,$$

which contradicts (47) with $k = 2$. Therefore, the singular control $u = 0$ is not optimal.

The necessary condition (47) with $k = 2$ was used in [37] to investigate the optimality of Lawden's spiral in the problem of optimizing the flight of a rocket in a central gravitational field in the free time case [38].

Remark 1. In deriving the necessary conditions in this section, we did not concretize the control variations $\delta u, \delta v, \cdots$. In [14] and [37] they are given in an explicit form on intervals of length 2τ, where $\tau > 0$ is small.

Remark 2. If we consider a needle variation to be the analogue of a δ-function, then the variations used in [14] and [37] can be represented as higher derivatives of δ-functions.

Remark 3. The sequence (47) of necessary conditions for optimality is also proved in [14] and [37], but in a different way on the basis of a special representation

for the second variation (35):

$$\delta^2 J = -\int_{t_0}^{t_1} \frac{\partial H(x+\delta x, \psi+\delta\psi, u+\delta u, t)}{\partial u}\delta u\,dt + \delta x'(t_1)\frac{\partial^2\phi}{\partial x^2}\delta x(t_1),$$

$$\delta\dot{\psi} = -\frac{\partial^2 H}{\partial x^2}\delta x - \frac{\partial^2 H}{\partial x\partial\psi}\delta\psi - \frac{\partial^2 H}{\partial u\partial x}\delta u, \qquad \delta\psi(t_1) = 0.$$

5. Optimality conditions for multidimensional singular controls. In the preceding sections, one-dimensional controls were considered. The general case was investigated in [39]–[42]. The methods described in these works are direct generalizations to the multidimensional case of the methods developed in [37] for one-dimensional singular controls (see § 4). It turns out [40] that, in the multidimensional case, the character of the optimality conditions for singular controls is determined by the rank of the matrix $R = \partial^2 H/\partial u_m \partial u_s$, with $m, s = 1, \cdots, r$. A feature which is peculiar to the multidimensional case is the appearance also of equality-type necessary optimality conditions.

5.1. We shall give a brief description of the method described in [39] and [40]. The second variation in the case of an r-dimensional control has the form

$$\delta^2 J = -\int_{t_0}^{t_1} (\delta u' R\,\delta u + 2\,\delta u' Q\,\delta x + \delta x' P\,\delta x)\,dt + \delta x'(t_1)\frac{\partial^2\phi(x(t_1))}{\partial x^2}\delta x(t_1), \tag{48}$$

where $R = \partial^2 H/\partial u^2$, $Q = \partial^2 H/\partial u\partial x$ and $P = \partial^2 H/\partial x^2$ are $r \times r$, $r \times n$, and $n \times n$ matrices, respectively; $\delta u = \{\delta u_1, \cdots, \delta u_r\}$, and

$$\delta\dot{x} = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial u}\delta u, \qquad \delta x(t_0) = 0. \tag{49}$$

A classical necessary condition for a minimum in problem (1)–(3) is the condition of Legendre and Clebsch,

$$\delta u' R\,\delta u = \sum_{m,s=1}^{r} \frac{\partial^2 H}{\partial u_m \partial u_s}\delta u_m\,\delta u_s \leqq 0, \tag{50}$$

obtained from the second variation (48). For a control which is singular in the classical sense (see Definition 2), the matrix R is singular, and the condition of Legendre and Clebsch becomes ineffective in the sense that (50) vanishes for nonzero variations δu. Following [37] (see the preceding section), it is possible also in this case to transform the second variation (48), under certain assumptions, with the aid of a change of variables of the type of (41). The condition of Legendre and Clebsch, derived from the transformed second variation, leads to a necessary optimality condition for singular controls.

Let the matrix R be transformed to the form

$$R = \begin{bmatrix} R_1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{51}$$

where R_1 is an $r^* \times r^*$ nonsingular matrix, and r^* is the rank of R. In accordance

with (51), let us decompose the matrices Q and $\partial f/\partial u$ as follows:

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}, \qquad \frac{\partial f}{\partial u} = [B_1, B_2],$$

where Q_1, Q_2, B_1 and B_2 are $r^* \times n$, $(r - r^*) \times n$, $n \times r^*$ and $n \times (r - r^*)$ matrices, respectively. Let us set $A = \partial f/\partial x$. We introduce the new variations

$$\begin{aligned} &\delta v_\alpha = \delta u_\alpha, \qquad \alpha = 1, \cdots, r^*, \\ &\delta v_\beta = \int_{t_0}^{t_1} \delta u_\beta \, d\tau, \qquad \beta = r^* + 1, \cdots, r, \\ &\delta y = \delta x - \sum_{\beta = r^*+1}^{r} \frac{\partial f}{\partial u_\beta} \delta v_\beta. \end{aligned} \tag{52}$$

It follows from (49) that δy and $\delta v = \{\delta v_\alpha, \delta v_\beta\}$ are related through the equation

$$\delta \dot{y} = \frac{\partial f}{\partial x} \delta y + \sum_{\alpha=1}^{r^*} \frac{\partial f}{\partial u_\alpha} \delta v_\alpha + \sum_{\beta=r^*+1}^{r} \left(\frac{\partial f}{\partial x} \frac{\partial f}{\partial u_\beta} - \frac{d}{dt} \frac{\partial f}{\partial u_\beta} \right) \delta v_\beta, \qquad \delta y(t_0) = 0. \tag{53}$$

Thus, the integral transformation is performed on the control variations δu_β which correspond to the linearly dependent (zero) rows of the matrix R. The new variations of the phase variables are introduced in such a way that (53) for these variations does not contain the variations δu_β with $\beta = r^* + 1, \cdots, r$.

Let

$$[Q_2 \quad B_2] = [Q_2 \quad B_2]'. \tag{54}$$

Then we can eliminate the old variables δu_β (see [40]) from (48). Thus, if (54) holds and $\delta v_\beta(t_1) = 0$ for $\beta = r^* + 1, \cdots, r$, then the second variation takes the form

$$\delta^2 J = -\int_{t_0}^{t_1} (\delta v' \bar{R}_2 \delta v + 2\delta v' \bar{Q} \delta y + \delta y' \bar{P} \delta y) \, dt + \delta y'(t_1) \frac{\partial^2 \phi}{\partial x^2} \delta y(t_1), \tag{55}$$

where

$$\begin{aligned} &\delta v = \{\delta v_1, \cdots, \delta v_{r^*}, \delta v_{r^*+1}, \cdots, \delta v_r\}, \\ &\bar{R}_2 = \begin{bmatrix} R_1 & R_2' \\ R_2 & R_3 \end{bmatrix}, \\ &R_2 = B_2' Q_1' - Q_2 B_1, \qquad R_3 = B_2' P B_2 - \frac{d}{dt} Q_2 B_2 - Q_2 B_3 - B_3 Q_2', \\ &B_3 = A B_2 - \dot{B}_2. \end{aligned} \tag{56}$$

The expressions for the matrices $\bar{Q}$ and $\bar{P}$ are given in [40].

It was proved in [40] that (54) is a necessary condition for optimality. Therefore, the form (55) of the second variation implies the following assertion.

Let the rank of the matrix $R = \partial^2 H/(\partial u_m \partial u_s)$ for a singular control $u(t)$, $t \in T$, be $r^* < r$, and let this matrix have the form (51). Then, in order that the control $u(t)$ be optimal, it is necessary that:

(a) the $(r - r^*) \times (r - r^*)$ matrix $Q_2 B_2$ be symmetric (an equality-type condition);

(b) if (a) holds, then the $r \times r$ matrix $\bar{R}_2$ must be negative semidefinite; i.e.,

$$\delta v' \bar{R}_2 \delta v \leqq 0 \quad \text{for every } \delta v. \tag{57}$$

Let us write down conditions (a) and (b) for the two-dimensional case ($r = 2$).

(i) Let the rank of R be zero. Then it is necessary that:

(a)
$$\frac{\partial^2 H'}{\partial u_1 \partial x}\frac{\partial f}{\partial u_2} = \frac{\partial^2 H'}{\partial u_2 \partial x}\frac{\partial f}{\partial u_1};$$

(b) if (a) holds, then

$$\sum_{i,j=1}^{r} \frac{\partial}{\partial u_i}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_j}\right)\eta_i \eta_j \geqq 0 \quad \text{for all } \eta_i, \eta_j. \tag{58}$$

In (58), a compact notation which follows from the representation (40) was used.

(ii) Let

$$R = \begin{bmatrix} \partial^2 H/\partial u_1^2 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \frac{\partial^2 H}{\partial u_1^2} \neq 0.$$

Then condition (a) becomes trivial, and condition (b) has the form

$$\frac{\partial^2 H}{\partial u_1^2}\eta_1^2 + 2\left(\frac{\partial^2 H'}{\partial u_1 \partial x}\frac{\partial f}{\partial u_2} - \frac{\partial^2 H'}{\partial u_2 \partial x}\frac{\partial f}{\partial u_1}\right)\eta_1 \eta_2 - \frac{\partial}{\partial u_2}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_2}\right)\eta_2^2 \leqq 0. \tag{59}$$

Example 6.

$$\dot{x}_1 = u_2, \quad \dot{x}_2 = x_1^2 + 4x_1 u_1 + u_1^2, \quad x_1(0) = x_2(0) = 0,$$

$$|u_1| \leqq 1, \quad |u_2| \leqq 1, \quad T = [0, 1], \quad \phi(x) = x_2,$$

$$H = \psi_1 u_2 - x_1^2 - 4x_1 u_1 - u_1^2, \quad \dot{\psi}_1 = 2x_1 + 4u_1, \quad \psi_1(1) = 0.$$

The control $u_1 = 0$, $u_2 = 0$ is singular on $[0, 1]$ according to Definition 2, since $\partial H/\partial u_1 = -4x_1 - 2u_1 = 0$, $\partial H/\partial u_2 = \psi_1 = 0$ and the matrix

$$R = \begin{bmatrix} -2 & 0 \\ 0 & 0 \end{bmatrix}$$

has rank 1. For this control,

$$\frac{\partial^2 H}{\partial u_1^2} = -2, \quad \frac{\partial^2 H'}{\partial u_1 \partial x}\frac{\partial f}{\partial u_2} - \frac{\partial^2 H'}{\partial u_2 \partial x}\frac{\partial f}{\partial u_1} = -4, \quad \frac{\partial}{\partial u_2}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_2}\right) = 2.$$

The quadratic form $-2\eta_1^2 - 8\eta_1\eta_2 - 2\eta_2^2$ is indefinite, which contradicts condition (59). The singular control under consideration cannot be optimal.

5.2. If a singular control is such that, in the matrix $\bar{R}_2$ (see (56)), the blocks $R_2 \equiv 0$ and $R_3 \equiv 0$, then $\bar{R}_2 = R$, and a single transformation is not effective.

In this case, a new transformation of the type of (52) is applied to (55), and the previously described procedure is repeated. As a result, we obtain a sequence of conditions [40]: In order that a singular control be optimal, it is necessary that:

(a) $Q_k B_k = [Q_k \quad B_k]'$;

(b) if (a) holds, then

$$\bar{R}_k = \begin{bmatrix} R_1 & R'_{2,k-2} \\ R_{2,k-2} & R_{3,k-2} \end{bmatrix} \leqq 0, \qquad k = 3, 4, \cdots.$$

Here, $Q_k B_k$ and $\bar{R}_k$ are $(r - r^*) \times (r - r^*)$ and $r \times r$ matrices, respectively, and

$$R_{2,k-2} = B'_k Q'_1 - Q_k B_1, \qquad R_{3,k-2} = B'_k P B_k - \frac{d}{dt} Q_k B_k - Q_k B_{k-1} - B'_{k-1} Q'_k,$$

$$Q_k = B'_{k-1} P - Q_{k-1} A - \dot{Q}_{k-1}, \qquad B_{k+1} = AB_k - \dot{B}_k.$$

The preceding necessary conditions were used in [40] to prove that a singular control in an interplanetary flight problem was not optimal [43].

5.3. Some of the results of § 5.1 and § 5.2 have been derived in a different form in [41] and [42] for systems of the form

$$\dot{x} = f_0(t, x, u) + \sum_{\mu=1}^{v} \alpha_\mu(t) f_{1\mu}(t, x, u), \tag{60}$$

where $\alpha = \{\alpha_1, \cdots, \alpha_v\}$ and $u = \{u_1, \cdots, u_r\}$ are controls which range over an open domain.

Let the rank of $[\partial^2 H/(\partial u_m \partial u_s)]$ be r. The control $\{\alpha, u\}$, for which

$$\partial H/\partial \alpha = 0, \qquad \partial H/\partial u = 0$$

is, in this case, singular according to Definition 2, since the rank of $[\partial^2 H/(\partial u_m \partial \alpha_\mu)]$ is equal to $r < r + v$.

To derive necessary conditions for optimality, an idea which was explained in the Appendix of [37], and which was based on a special representation of the second variation (see Remark 3, § 4), was used in [41] and [42]. The results are as follows: In order that a singular control of system (60) be optimal, it is necessary that:

(a) the equality-type conditions

$$\frac{\partial}{\partial \alpha_\mu}\left(\frac{d}{dt}\frac{\partial H}{\partial \alpha_p}\right) = 0, \qquad \mu, p = 1, 2, \cdots, v, \quad \mu > p, \tag{61}$$

hold;

(b) if (a) holds, then

$$\sum_{\mu,p=1}^{v} \frac{\partial}{\partial \alpha_\mu}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial \alpha_p}\right)\delta\alpha_\mu \delta\alpha_p + 2\sum_{\mu=1}^{v}\sum_{m=1}^{r} \frac{\partial}{\partial u_m}\left(\frac{d}{dt}\frac{\partial H}{\partial \alpha_\mu}\right)\delta u_m \delta\alpha_\mu$$

$$- \sum_{m,s=1}^{r} \frac{\partial^2 H}{\partial u_m \partial u_s}\delta u_m \delta u_s \geqq 0 \tag{62}$$

for all $\delta\alpha = \{\delta\alpha_1, \cdots, \delta\alpha_v\}$ and $\delta u = \{\delta u_1, \cdots, \delta u_r\}$.

It is not difficult to see that conditions (61) and (62) are a particular case of conditions (54) and (57), which were obtained earlier in [40]. Unlike (54) and (57), conditions (61) and (62) are expressed directly in terms of the Hamiltonian of the system. Further, they are closely related to the methods of evaluating singular controls (see § 2), and they are of a form quite similar to (40) for the one-dimensional case.

6. The method of transformations in state space. In the preceding sections, singular controls were investigated with the aid of the second variation of the functional being minimized. This variation was either evaluated for a specially chosen control variation (§ 3), or was transformed, with regard to a certain change of variations, to a form analogous to the nonsingular case, which made it possible to apply effectively the conditions of Legendre and Clebsch (§§ 4 and 5). Another approach to an analysis of the controls is the method of transformations in state space, which was proposed in [44] and was developed in [45]–[48]. The method is based on a change of phase variables, which leads to a new problem of smaller dimension. In the latter, singular controls of the original problem can be investigated, generally speaking, by classical methods of the calculus of variations. As is well known, these methods are ineffective for singular controls in the original problem. An essential difficulty for a practical application of this method is the necessity of having to find independent first integrals of a nonlinear system of ordinary differential equations. However, in many applied problems, this difficulty can be overcome (see, for example, [14]).

6.1. Following [44], we shall describe the method of transformations in a state space for the one-dimensional control case.

We are given the system

$$\begin{aligned} &\dot{x}_i = f_{0i}(x_1, \cdots, x_n, t) + u f_{1i}(x_1, \cdots, x_n, t), \qquad i = 1, \cdots, n, \\ &u(t) \in U, \quad t \in T, \quad J(u) = \phi(x(t_1)). \end{aligned} \tag{63}$$

Let us consider new phase variables z_j according to the formulas

$$z_j = r_j(x_1, \cdots, x_n, t), \qquad j = 1, \cdots, n-1.$$

We choose the functions r_j such that the equations for the new variables

$$\dot{z}_j = \sum_{i=1}^{n} \frac{\partial r_j}{\partial x_i} f_{0i} + \sum_{i=1}^{n} \frac{\partial r_j}{\partial x_i} f_{1i} \cdot u + \frac{\partial r_j}{\partial t} \tag{64}$$

do not contain the control u. To do this, it is sufficient to set

$$\sum_{i=1}^{n} \frac{\partial r_j}{\partial x_i} f_{1i} = 0, \qquad j = 1, \cdots, n-1. \tag{65}$$

Solving the partial differential equation (65) by the method of characteristics [49], we arrive at the following system of ordinary differential equations:

$$\frac{dx_i}{dt} = f_{1i}(x_1, \cdots, x_n, t), \qquad i = 1, \cdots, n.$$

Let $c_k = \phi_k(x_1, \cdots, x_n, t)$, $k = 1, \cdots, n-1$, $c_k = \text{const.}$, be independent first integrals of this system. Each of these integrals is a solution of (65). Therefore, we can set

$$z_j = r_j = \phi_j(x_1, \cdots, x_n, t), \quad j = 1, \cdots, n-1, \qquad z_n = x_l, \tag{66}$$

where l is such that

$$f_{1l}(x, t) \neq 0. \tag{67}$$

If we apply the nonsingular transformation (66), then we obtain from (64) the following equations for the new variables:

$$\begin{aligned} \dot{z}_j &= b_j(z_1, \cdots, z_n, t), & j &= 1, \cdots, n-1, \\ \dot{z}_n &= a_n(z_1, \cdots, z_n, t) + b_n(z_1, \cdots, z_n, t)u. \end{aligned} \tag{68}$$

Now a singular control is characterized by the identity

$$\partial H/\partial u = \psi_n b_n(z_1, \cdots, z_n, t) \equiv 0, \tag{69}$$

where $b_n(z_1, \cdots, z_n, t) \not\equiv 0$ by virtue of (67).

Thus, for a singular control,

$$\psi_n(t) \equiv 0, \qquad \dot{\psi}_n = -\partial H/\partial z_n \equiv 0. \tag{70}$$

If we now consider the system of $n-1$ equations

$$\dot{z}_j = b_j(z_1, \cdots, z_n, t), \qquad j = 1, 2, \cdots, n-1, \tag{71}$$

where $z_1, \cdots, z_{n-1}$ are phase variables and z_n is the control, then, by virtue of (70), the Hamiltonians of systems (68) and (71) coincide, and condition (69), expressing the singularity of the control u in (68), corresponds to the condition that the Hamiltonian of system (71) is stationary with respect to z_n. The end purpose of the method is a changeover from the n-dimensional system (63) to the $(n-1)$-dimensional system (71). In the optimization problem in the new variables, the classical conditions (in particular, the condition $\partial^2 H/\partial z_n^2 \leqq 0$ of Legendre and Clebsch) may turn out to be effective. If the control z_n enters (71) linearly, the described transformation is repeated.

6.2. Another interpretation of the method of transformations has been given in [45]–[48]. In [45], as opposed to [44], the connection between the original problem (for system (63)) and the transformed one (for system (71)) was studied in more detail. It turned out that it is automatically possible to construct a solution of the original problem, using a solution of the transformed one ("Problem 2" in the terminology of [46]), if (i) $-\infty \leqq u \leqq \infty$, (ii) $u_0 \leqq u \leqq \infty$, and if certain additional conditions are satisfied. In the general case, the original and the transformed problems are not equivalent [45]. In investigating case (ii) in [48], a new type of sliding regime ("cyclical sliding regime") was discovered. The works [50] and [51] were devoted to the problem of the realization of such regimes. It was shown in [47] that, with the aid of the method of transformations, the degenerate conjugate problem of a minimum of the second variation [3] can be effectively solved. This provides additional information about the character of a

singular control in the original problem. In [46], the method of transformations was considered for systems which are nonlinear in the control.

By means of the method of transformations in state space, a number of flight dynamic problems have been solved: The problem of optimal trajectories of a jet aircraft in a central field [52], the problem of optimal transfers between coplanar elliptical orbits in a central field [53], the problem of the optimality of singular regimes of motions of rockets in a central field [54] (the Lawden spiral problem), the problem of the dynamics of a point with variable mass in a homogeneous gravitational field [55], and others.

7. Bundle of control variations.

7.1. We shall describe one more approach to investigating singular controls which is based on a new type of control variations [56], [57]. By definition, a variation $\delta u(t)$ will be said to be a bundle (of first order) at a point $\theta \in T$ if

$$\delta u(t) = \sum_{i=1}^{p} \delta u^i(t), \qquad t \in T,$$

$$\delta u^i(t) = \begin{cases} \phi_i(t), & \theta \leqq t < \theta + \varepsilon_i, \\ 0, & t \notin [\theta, \theta + \varepsilon_i), \end{cases}$$

where $\phi_i(t), \theta \leqq t < \theta + \varepsilon_i$, are arbitrary functions, $\varepsilon_i = q_i\varepsilon, \varepsilon > 0, i = 1, 2, \cdots, p$, $0 < q_j < q_{j+1} \leqq 1, j = 1, 2, \cdots, p-1$ and $q_p = 1$.

Higher order bundles of variations are defined in [57].

As one can see from the definition, a bundle of variations is an immediate generalization of a "needle variation." A bundle is defined on an interval of length ε and is characterized by the parameters $\phi_1, \cdots, \phi_p$, and $q_1, \cdots, q_p$. All of the known variations which were used in [34], [37], [41] and [42] may be extracted from a bundle. The general idea of how to obtain necessary conditions for optimality with the aid of a bundle of variations consists of the following.

An increment $\Delta J(u)$ in the functional (3) which is generated by a bundle may be expanded in powers of ε:

$$\Delta J(u) = \sum_{z=1}^{s} R_z \varepsilon^z + o(\varepsilon^s). \tag{72}$$

Since the coefficients of ε^z in (72) is a function of the bundle parameters $R_z = R_z(\phi_1, \cdots, \phi_p, q_1, \cdots, q_p)$, one can, by changing the structure of a bundle, first of all obtain all of the possible simplified optimality criteria from the coefficient R_z under consideration, and, second of all, go over to an investigation of the next coefficient R_{z+1} by choosing a bundle such that $R_z(\phi_1, \cdots, \phi_p, q_1, \cdots, q_p) = 0$. If, in investigating the principal term in (72) (of order ε^z) associated with a bundle, the coefficient of ε^z has constant sign on the bundle (i.e., if its sign does not depend on the bundle parameters), then inequality-type necessary conditions are obtained from $\Delta J(u^0) \geqq 0$. In the opposite case (if the sign of the coefficient changes on the bundle), equality-type conditions are obtained.

7.2. Let us single out necessary conditions for optimality of a one-dimensional singular control, which are to be derived by investigating the second variation for

a first order bundle. Let the bundle consist of constant functions $\phi_i(t) = \phi_i$, $i = 1, \cdots, p$. The expansion of the second variation on this bundle has the following form, to within terms of order $o(\varepsilon^4)$ (see [57]):

$$
\begin{aligned}
-\delta^2 J = N(\theta)\left(\sum_{i=1}^{p} \phi_i q_i\right)^2 \varepsilon^2 + \left\{\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) M(\phi_1, \cdots, \phi_p, q_1, \cdots, q_p)\right. \\
\left. + \frac{1}{2}\frac{dN(t)}{dt}\left(\sum_{i=1}^{p} \phi_i q_i\right)\left(\sum_{i=1}^{p} \phi_i q_i^2\right)\right\}_{t=\theta} \varepsilon^3 \\
+ \left\{\frac{d}{dt}\left[\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right)\right] M(\phi_i, \cdots, \phi_p, q_1, \cdots, q_p) + \frac{1}{8}\frac{d^2 N(t)}{dt^2}\left(\sum_{i=1}^{p} \phi_i q_i^2\right)^2\right. \\
\left. + N_1(t)\left[\frac{1}{3}\left(\sum_{i=1}^{p} \phi_i q_i\right)\left(\sum_{i=1}^{p} \phi_i q_i^3\right) - \frac{1}{4}\left(\sum_{i=1}^{p} \phi_i q_i^2\right)^2\right]\right\}_{t=\theta} \varepsilon^4 + o(\varepsilon^4).
\end{aligned} \tag{73}
$$

Here,

$$
\begin{aligned}
N(t) &= \frac{\partial^2 H'}{\partial u \partial x}\frac{\partial f}{\partial u} + \frac{\partial f'}{\partial u}\Psi\frac{\partial f}{\partial u}, \\
N_1(t) &= \frac{\partial^2 H'}{\partial u \partial x}\left(\frac{d^2}{dt^2}\frac{\partial f}{\partial u} - \frac{d}{dt}\frac{\partial f}{\partial x}\cdot\frac{\partial f}{\partial u} - 2\frac{\partial f}{\partial x}\frac{d}{dt}\frac{\partial f}{\partial u} + \frac{\partial f}{\partial x}\frac{\partial f}{\partial x}\frac{\partial f}{\partial u}\right) \\
&\quad + \frac{\partial f'}{\partial u}\Psi\left(\frac{d^2}{dt^2}\frac{\partial f}{\partial u} - \frac{d}{dt}\frac{\partial f}{\partial x}\cdot\frac{\partial f}{\partial u} - 2\frac{\partial f}{\partial x}\frac{d}{dt}\frac{\partial f}{\partial u} + \frac{\partial f}{\partial x}\frac{\partial f}{\partial x}\frac{\partial f}{\partial u}\right), \\
M(\phi_1, q_1) &= \frac{1}{6}\phi_1^2 q_1^3, \\
M(\phi_1, \phi_2, q_1, q_2) &= \frac{2}{3}(\phi_1^2 q_1^3 + \phi_2^2 q_2^3) + \frac{1}{3}\phi_1\phi_2 q_1^3 + \phi_2\phi_1 q_2^2 q_1 \\
&\quad - \frac{1}{2}(\phi_1 q_1 + \phi_2 q_2)(\phi_1 q_1^2 + \phi_2 q_2^2), \\
M_1(\phi_1, q_1) &= \frac{1}{8}\phi_1^2 q_1^4,
\end{aligned}
$$

where the $n \times n$ matrix function Ψ satisfies the equation

$$
\dot{\Psi} = -\frac{\partial f'}{\partial x}\Psi - \Psi\frac{\partial f}{\partial x} - \frac{\partial^2 H}{\partial x^2}, \quad \Psi(t_1) = -\frac{\partial^2 \phi}{\partial x^2}. \tag{74}
$$

Let us note that all of the derivatives with respect to time in (73) are from the right by virtue of the fact that the variations $\delta u(t)$ are given explicitly on $[\theta, \theta + \varepsilon)$. A realization of the general idea of § 7.1 in the particular case under consideration leads to the following necessary conditions for optimality of singular controls:

(a) Let a bundle be arbitrary. Then, from the form of the coefficient of ε^2 in (73), it follows that

$$
N(t) \leqq 0, \qquad t \in T. \tag{75}
$$

An illustration. None of the conditions (47) is effective for the singular control $u = 0$ of Example 4, while (75) turns out to be effective:

$$\frac{\partial^2 H'}{\partial u\, \partial x} \frac{\partial f}{\partial u} + \frac{\partial f'}{\partial u} \Psi \frac{\partial f}{\partial u} = 1 > 0,$$

i.e., the singular control under consideration is not optimal.

(b) Let us impose on a bundle the condition

$$\sum_{i=1}^{p} \phi_i q_i = 0. \tag{76}$$

The principal term in (73), under condition (76), has the form ($p = 2$)

$$-\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right)\Bigg|_{t=\theta} \frac{q_2(q_2 - q_1)^2}{3}\phi_2^2 \cdot \varepsilon^3.$$

Since the coefficient $(q_2(q_2 - q_1)^2/3)\phi_2^2 \geqq 0$, the necessary condition for optimality (40) follows.

(c) Now let a singular control be such that

$$N(\theta) = 0, \qquad \theta \in T. \tag{77}$$

Setting $p = 1$ in (73), we obtain

$$\left[\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) + 3\frac{dN(t)}{dt}\right]_{t=\theta} \leqq 0, \qquad \theta \neq t_1. \tag{78}$$

If the bundle is defined on $(\theta - \varepsilon, \theta]$, we obtain the condition

$$\left[\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) - 3\frac{dN(t)}{dt}\right]_{t=\theta} \leqq 0, \qquad \theta \neq t_0, \tag{79}$$

where d/dt denotes the derivative from the left.

An illustration. For the singular control $u = 0$ of Example 3, $N(t_1) = N(1) = 0$. The left-hand inequality in (79) has the form

$$\left[\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) - 3\frac{dN(t)}{dt}\right]_{t=1} = 4,$$

which contradicts (79). The singular control $u = 0$ is not optimal.

Let us now investigate the coefficient of ε^4.

(d) Let us choose a bundle which satisfies condition (76), and let a singular control be such that

$$\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right)\Bigg|_{t=\theta} = \frac{d}{dt}\left[\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right)\right]_{t=\theta} = 0. \tag{80}$$

Expanding (73), we obtain the necessary condition for optimality

$$d^2N(t)/dt^2|_{t=\theta} - 2N_1(\theta) \leqq 0. \tag{81}$$

An illustration. Equations (80) are satisfied identically for the singular control $u = 0$ of Example 5. Let us apply condition (81):

$$d^2N(t)/dt^2 - 2N_1(t) = -4(t-1) \geqq 0, \qquad t \in [0,1].$$

The singular control is not optimal.

(e) Let a singular control be such that (77) holds and (78) holds as an equality. Then, from (73) with $p = 1$, we have

$$\left\{\frac{d}{dt}\left[\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right)\right] + \frac{d^2N(t)}{dt^2}\right\}_{t=\theta} + \tfrac{2}{3}N_1(\theta) \leqq 0, \qquad \theta \neq t_1. \tag{82}$$

It follows from the results of § 7.1 that the necessary conditions for optimality of singular controls (40) and (75) are unconditional. These criteria are related in the following way [58]: If, for a singular optimal control,

$$N(\theta) = dN/dt|_{t=\theta} = 0,$$

then also

$$\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right)\Bigg|_{t=\theta} = 0.$$

In other words, if condition (75) is ineffective for a singular optimal control, then so is condition (40). The converse, generally speaking, is not true [58]. We obtain from this the following criterion for nonoptimality: If, for a singular control, condition (75) holds as an equality, and if the left-hand side of (40) is nonzero, then the singular control being tested is not optimal.

Example 7.

$$\dot{x}_1 = u, \quad \dot{x}_2 = x_1 + x_2, \quad \dot{x}_3 = -ux_2, \quad x_1(0) = x_2(0) = x_3(0) = 0,$$
$$T = [0,1], \quad |u| < 1, \quad \phi(x) = x_3.$$

For the singular control $u \equiv 0$,

$$\frac{\partial^2 H'}{\partial u \partial x}\frac{\partial f}{\partial u} + \frac{\partial f'}{\partial u}\Psi\frac{\partial f}{\partial u} \equiv 0, \qquad \frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) = 2 > 0.$$

Although condition (40) is satisfied, the control $u \equiv 0$ is not optimal.

7.3. The two-dimensional control case. Let us consider problem (1)–(3), where $u = \{u_1, u_2\}$. Let us assign a control variation in the form of a first order bundle:

$$\delta u_1(t) = \sum_{i=1}^{p} \delta u_1^i(t), \qquad \delta u_1^i(t) = \begin{cases} a_i, & \theta \leqq t < \theta + q_i\varepsilon, \\ 0, & t \notin [\theta, \theta + q_i\varepsilon), \end{cases}$$
$$0 < q_i < q_{i+1} \leqq 1, \quad q_p = 1;$$

$$\delta u_2(t) = \sum_{j=1}^{q} \delta u_2^j(t), \qquad \delta u_2^j(t) = \begin{cases} b_j, & \theta \leqq t < \theta + p_j\varepsilon, \\ 0, & t \notin [\theta, \theta + p_j\varepsilon), \end{cases}$$
$$0 < p_j < p_{j+1} \leqq 1, \quad p_q = 1.$$

The matrix $R = [\partial^2 H/(\partial u_m \partial u_s)]$, $m, s = 1, 2$, becomes singular for a singular control. There are two possible cases: (i) rank $R = 0$, and (ii) rank $R = 1$.

(i) Rank $R = 0$.

(a) For a bundle with parameters $p = 1$, $q = 1$, $a_1 = \eta_1$, and $b_1 = \eta_2$, the expansion of the second variation has the form

$$-\delta^2 J = N(\theta, \eta_1, \eta_2)\varepsilon^2 + \frac{1}{6}\left[L(\theta, \eta_1, \eta_2) + 3\frac{d}{dt}N(\theta, \eta_1, \eta_2)\right]\varepsilon^3 + o(\varepsilon^3), \tag{83}$$

where

$$N(t, \eta_1, \eta_2) = \sum_{i,j=1}^{2} \left(\frac{\partial^2 H'}{\partial u_i \partial x}\frac{\partial f}{\partial u_j} + \frac{\partial f'}{\partial u_i}\Psi\frac{\partial f}{\partial u_j}\right)\eta_i\eta_j,$$

$$\frac{d}{dt}N(t, \eta_1, \eta_2) = \sum_{i,j=1}^{2} \frac{d}{dt}\left(\frac{\partial^2 H'}{\partial u_i \partial x}\frac{\partial f}{\partial u_j} + \frac{\partial f'}{\partial u_i}\Psi\frac{\partial f}{\partial u_j}\right)\eta_i\eta_j,$$

$$L(t, \eta_1, \eta_2) = \sum_{i,j=1}^{2} \frac{\partial}{\partial u_i}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_j}\right)\eta_i\eta_j,$$

and the matrix Ψ satisfies (74). It follows from (83) that the quadratic form N is negative semidefinite:

$$N(t, \eta_1, \eta_2) \leqq 0, \qquad t \in T. \tag{84}$$

(b) If, for the same bundle and for some $t = \theta$,

$$N(\theta, \eta_1, \eta_2) = 0,$$

then we have the following optimality condition:

$$L(\theta, \eta_1, \eta_2) + 3\frac{d}{dt}N(\theta, \eta_1, \eta_2) \leqq 0.$$

(c) Let us consider bundles with $p = q = 2$ and with parameters such that

$$\sum_{i=1}^{2} a_i q_i = 0, \qquad \sum_{i=1}^{2} b_i p_i = 0.$$

The expansion of the second variation in this case has the form

$$\begin{aligned}
-\delta^2 J = &\left[\left(\frac{\partial^2 H'}{\partial u_1 \partial x}\frac{\partial f}{\partial u_2} - \frac{\partial^2 H}{\partial u_2 \partial x}\frac{\partial f}{\partial u_1}\right)a_1 b_1 p_1(1 - q_1)(q_1 - p_1)\right]_{t=\theta}\varepsilon^2 \\
&- \frac{1}{3}\left\{\frac{\partial}{\partial u_1}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_1}\right)a_1^2 q_1^2(1 - q_1)^2 + \frac{1}{2}\left[\frac{\partial}{\partial u_1}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_2}\right)\right.\right. \\
&\left. + \frac{\partial}{\partial u_2}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_1}\right)\right]a_1 b_1 p_1(1 - q_1)(2q_1 - p_1^2 - q_1^2) \\
&+ \frac{\partial}{\partial u_2}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_2}\right)b_1^2 p_1^2(1 - p_1)^2 \\
&\left. - \frac{3}{2}\frac{d}{dt}\left(\frac{\partial^2 H'}{\partial u_1 \partial x}\frac{\partial f}{\partial u_2} - \frac{\partial^2 H'}{\partial u_2 \partial x}\frac{\partial f}{\partial u_1}\right)a_1 b_1 p_1(1 - q_1)(q_1^2 - p_1^2)\right\}_{t=\theta}\varepsilon^3 + o(\varepsilon^3).
\end{aligned} \tag{85}$$

The coefficient of ε^2 in (85) may be assigned either a positive or a negative sign by the choice of the bundle parameters a_1 and b_1. Therefore, for a singular control to be optimal, it is necessary that the equality

$$\frac{\partial^2 H'}{\partial u_1 \partial x}\frac{\partial f}{\partial u_2} - \frac{\partial^2 H'}{\partial u_2 \partial x}\frac{\partial f}{\partial u_1} = 0 \tag{86}$$

hold. This equality can be represented in the form

$$\frac{\partial}{\partial u_2}\left(\frac{d}{dt}\frac{\partial H}{\partial u_1}\right) = 0.$$

(d) Let us consider the bundles of (c) with $p_1 = q_1$. Then it follows from (85) that

$$L(t, \eta_1, \eta_2) \geqq 0, \qquad t \in T. \tag{87}$$

Remark 1. Condition (87) has been obtained without the assumption (86). In [40] and [41], in a similar situation, the inequality-type criterion (87) was obtained, providing the equality-type criterion (86) held.

Remark 2. If, for a singular control, (86) holds, then the matrix of the quadratic form (87) is symmetric:

$$\frac{\partial}{\partial u_1}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_2}\right) = \frac{\partial}{\partial u_2}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_1}\right).$$

Remark 3. As in the one-dimensional case, there is a relation between (84) and (87): If, for some θ, η_1, and η_2,

$$N(\theta, \eta_1, \eta_2) = \frac{d}{dt}N(\theta, \eta_1, \eta_2) = 0,$$

then

$$L(\theta, \eta_1, \eta_2) = 0.$$

(ii) Let rank $R = 1$. Set

$$Q = \left\{\eta_1, \eta_2 : \sum_{i,j=1}^{2} \frac{\partial^2 H}{\partial u_i u_j}\eta_1 \eta_2 = 0\right\}.$$

(a) Let us consider bundles with parameters $p = q = 1, a_1 = \eta_1$ and $b_1 = \eta_2$. In this case, the second variation has the form (83) on the elements of the set Q. Therefore, it follows from a consideration of the elements of order ε^2 in (83) that

$$N(t, \eta_1, \eta_2) \leqq 0, \qquad t \in T, \quad \eta_1, \eta_2 \in Q.$$

(b) Set $N(t, \eta_1, \eta_2) = 0$, $t \in T$, $\eta_1, \eta_2 \in Q$. If we consider the terms of order ε^3 in (83) for the same bundles, we are led to the following necessary condition for optimality:

$$L(t, \eta_1, \eta_2) + 3\frac{d}{dt}N(t, \eta_1, \eta_2) \leqq 0, \qquad t \in T, \quad \eta_1, \eta_2 \in Q.$$

(c) Let the bundle parameters satisfy the conditions

$$p = q = 2, \quad \sum_{i=1}^{2} a_i q_i = 0, \quad \sum_{i=1}^{2} b_i p_i = 0, \quad p_1 = q_1, \quad a_1 = \eta_1, \quad b_1 = \eta_2.$$

Let us write out the second variation for the elements $\eta_1, \eta_2 \in Q$:

$$-\delta^2 J = \frac{1}{3}\left\{\frac{3}{2}\left(\frac{d^2}{dt^2}\frac{\partial^2 H}{\partial u_1^2}\eta_1^2 + 2\frac{d^2}{dt^2}\frac{\partial^2 H}{\partial u_1 \partial u_2}\eta_1\eta_2 + \frac{d^2}{dt^2}\frac{\partial^2 H}{\partial u_2^2}\eta_2^2\right)\right.$$
$$-\left[\frac{\partial}{\partial u_1}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_1}\right)\eta_1^2 + \frac{\partial}{\partial u_1}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_2}\right)\eta_1\eta_2\right.$$
$$\left.+\frac{\partial}{\partial u_2}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_1}\right)\eta_1\eta_2 + \frac{\partial}{\partial u_2}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_2}\right)\eta_2^2\right]$$
$$\left.\cdot(1 - q_1)\right\}_{t=\theta} q_1^2(1 - q_1)\varepsilon^3 + o(\varepsilon^3).$$

Separating out the principal terms with respect to q_1, $0 < q_1 < 1$, in the coefficient of ε^3, we obtain the following optimality conditions:

$$\sum_{i,j=1}^{2} \frac{d^2}{dt^2}\frac{\partial^2 H}{\partial u_i \partial u_j}\eta_i\eta_j \leqq 0, \qquad \eta_1, \eta_2 \in Q,$$
$$\sum_{i,j=1}^{2}\left[\frac{3}{2}\frac{d^2}{dt^2}\frac{\partial^2 H}{\partial u_i \partial u_j} - \frac{\partial}{\partial u_i}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_j}\right)\right]\eta_i\eta_j \leqq 0, \qquad \eta_1, \eta_2 \in Q.$$

Remark 4. The equality-type condition (86) and the generalized condition of Kelley (87), when rank $R = 1$, may not be satisfied by an optimal control.

Example 8.

$$\dot{x}_1 = u_1 x_2 - u_2, \quad \dot{x}_2 = u_1, \quad \dot{x}_3 = x_1^2 + (u_1 x_2 - u_2)^2,$$
$$x_1(0) = x_2(0) = x_3(0) = 0, \quad T = [0, 1], \quad |u_1| < 2, \quad |u_2| < 2, \quad \phi(x) = x_3.$$

The control $u_1(t) \equiv 1, u_2(t) = t$ is a singular optimal control. For this control, condition (86)

$$\frac{\partial^2 H'}{\partial u_1 \partial x}\frac{\partial f}{\partial u_2} - \frac{\partial^2 H'}{\partial u_2 \partial x}\frac{\partial f}{\partial u_1} = -2 \neq 0$$

and condition (87)

$$L(t, \eta_1, \eta_2) = -4\eta_2^2/t^2 \leqq 0$$

(with $\eta_1, \eta_2 \in Q$) do not hold.

Remark 5. A generalization of the described method to the r-dimensional control case ($r > 2$) does not cause any basic difficulties.

7.4. In problem (1)–(3), let the control region U be a closed body with a boundary which is, in a neighborhood of an optimal control, a smooth $(r - 1)$-dimensional manifold. Let a denote the normal to the tangent plane to U at the point $u(t)$. Let rank $R = 0$. Let us derive the formula for the increment of a

functional on a first order bundle. Making use of the particular structure of a bundle, it is possible to show that, for a control $u(t)$, which is singular in the classical sense and which lies on the boundary of U, to be optimal, it is necessary that the equations

$$\frac{\partial^2 H'}{\partial u_i \partial x}\frac{\partial f}{\partial u_j} - \frac{\partial^2 H'}{\partial u_j \partial x}\frac{\partial f}{\partial u_i} = 0, \qquad i,j = 1,2,\cdots,r, \quad i<j, \tag{88}$$

hold.

The necessary conditions for optimality (88) are equality-type conditions which coincide with the corresponding conditions for singular controls that lie interior to the admissible region U (see (61) and (86)).

A further investigation of the increment of the cost functional on a bundle shows that, for an optimal control which is singular in the classical sense and which lies on the boundary of U, the following condition must also hold:

$$\sum_{i,j=1}^{r} \frac{\partial}{\partial u_i}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_j}\right)\eta_i\eta_j \geqq 0 \tag{89}$$

for all $\eta = (\eta_1, \cdots, \eta_r)$ such that

$$a'\eta = 0. \tag{90}$$

The optimality condition (89), (90) is similar to condition (87) for an open region U, but it contains the additional restriction (90). It is not difficult to show that, in the general case, it is not possible to omit the constraint (90).

Example 9.

$$\dot{x}_1 = u_1, \quad \dot{x}_2 = u_2 - 1, \quad \dot{x}_3 = u_2 + 1, \quad \dot{x}_4 = x_1^2 + (u_2 - 1)x_2x_3,$$

$$x_1(0) = x_2(0) = x_3(0) = x_4(0) = 0,$$

$$U = \{u_1, u_2 : |u_1| \leqq 1, |u_2| \leqq 1\}, \quad T = [0,1], \quad \phi(x) = x_4.$$

The control $u_1 \equiv 0$, $u_2 \equiv 1$ is a singular optimal one, and it lies on the boundary of U. For this control, the quadratic form

$$\sum_{i,j=1}^{2} \frac{\partial}{\partial u_i}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u_j}\right)\eta_i\eta_j = 2\eta_1^2 - 2\eta_2^2 \tag{91}$$

is indefinite, which contradicts (87). But, under condition (90), the form (91) is positive definite, which does not contradict (89), (90).

8. Necessary conditions for optimality for closed control domains. Matrix impulses.

8.1. It is well known that the maximum principle arose in connection with the appearance of optimization problems in which the control terms were to be chosen from closed sets. In such cases, the optimal controls typically lie on the boundary of the admissible sets. As follows from the preceding section, the necessary conditions for optimality derived under the assumption that U is open turn out, generally speaking, to be weaker when we consider the case of U closed. Moreover, as was shown in [59], the optimality conditions obtained in [34], [37] and [40]–[42], and presented in the preceding sections, cannot, in the

general case, be generalized to singular controls which lie on the boundary of U. The reason for this is hidden in the peculiarities of the variations used to derive these conditions. Essentially, the variations in [34], [37] and [40]–[42] are two-sided variations. Therefore, they cannot, generally speaking, be applied to problems in which U is closed.

Let us show, for example, following [59], that the necessary condition for optimality (40) becomes false for singular controls which lie on the boundary of U.

Example 10.

$$\dot{x}_1 = u - 1, \quad \dot{x}_2 = u + 1, \quad \dot{x}_3 = (u - 1)x_1x_2, \quad x_1(0) = x_2(0) = x_3(0) = 0,$$

$$U = \{u : |u| \leqq 1\}, \quad T = [0, 1], \quad \phi(x) = x_3.$$

The boundary control $u(t) \equiv 1$ is here both optimal and singular on T (by Definition 2 of § 1). It is not difficult to make the evaluation that

$$\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) = -2,$$

which contradicts (40).

Necessary conditions for optimality of singular (in the sense of the maximum principle) controls when U is closed were worked out in [59]–[61]. Matrix impulses were used to formulate these conditions. Some elements of such an approach were described in the preceding section for the problem with an open control set. All of the considerations in this section are carried out for an arbitrary region U.

8.2. Let the problem under consideration be problem (1)–(3) in which U is a closed set. The increment of the functional (3) on the two admissible controls $u(t)$ and $\tilde{u}(t) = u(t) + \Delta u(t)$, and the corresponding trajectories $x(t)$ and $\tilde{x}(t) = x(t) + \Delta x(t)$ can be represented by the form [61]

$$\begin{aligned}\Delta J(u) = &-\int_{t_0}^{t_1} \Delta_{\tilde{u}} H(x(t), \psi(t), u(t), t)\, dt \\ &-\int_{t_0}^{t_1} \left\{\frac{\partial \Delta_{\tilde{u}} H'(x(t), \psi(t), u(t), t)}{\partial x}\right\} \Delta x(t)\, dt \\ &-\int_{t_0}^{t_1} \Delta x'(t)\Psi(t)\Delta_{\tilde{u}} f(x(t), u(t), t)\, dt + \eta,\end{aligned} \tag{92}$$

where ψ is a vector impulse which satisfies (5), Ψ is a matrix impulse evaluated by virtue of (74), and η is the remainder term.

Let us assume that $\tilde{u}(t)$ is generated by the needle variation

$$\tilde{u}(t) = \begin{cases} u(t), & t \notin [\theta, \theta + \varepsilon), \\ v, & t \in [\theta, \theta + \varepsilon). \end{cases} \tag{93}$$

Separating out the principle term with respect to ε in the increment (92) caused by the control (93), we obtain, for a control singular in the sense of the

maximum principle (Definition 1 of § 1), the following necessary condition for optimality:

$$\left\{\frac{\partial \Delta_v H(x(t), \psi(t), u(t), t)}{\partial x}\right\}' \Delta_v f(x(t), u(t), t) + \Delta_v f'(x(t), u(t), t)\Psi(t)\Delta_v f(x(t), u(t), t) \leqq 0, \qquad v \in \omega(t), \quad t \in T. \tag{94}$$

Hence, it follows that, for convex U,

$$\left\{\frac{\partial^2 H(x(t), \psi(t), u(t), t)}{\partial x \partial u}[v - u(t)]\right\}' \frac{\partial f(x(t), u(t), t)}{\partial u}[v - u(t)] + \left\{\frac{\partial f(x(t), u(t), t)}{\partial u}[v - u(t)]\right\}' \Psi(t) \frac{\partial f(x(t), u(t), t)}{\partial u}[v - u(t)] \leqq 0,$$

$$v \in \omega(t), \quad t \in T.$$

If U is an open set (or if an optimal control $u^0(t)$ passes through interior points of U), then the necessary condition for optimality (84) is obtained from (94).

Let us give an application of the necessary condition (94) for the singular control of Example 10. It is not difficult to check that, in this case, all of the elements of the matrix $\Psi(t)$ are zero: $\psi_{ij}(t) \equiv 0$, $i, j = 1, 2, 3$. Condition (94) takes the following form for the control $u(t) \equiv 1$:

$$\frac{\partial \Delta_v H'}{\partial x} \Delta_v f = -2t(v - 1)^2 \leqq 0, \qquad |v| \leqq 1, \quad t \in [0, 1];$$

i.e., the singular control under consideration is a candidate for an optimal control.

8.3. The necessary conditions for optimality which we have presented are second order conditions. Higher order criteria are discussed in [61], where formulas for increments of arbitrary order are derived. The method of proof, which is based on these formulas, allows one to restrict oneself to needle variations. By virtue of this fact, the domain of application of the results includes arbitrary bounded control sets U. Additional assumptions on U lead to new results. In [62], high order necessary conditions for optimality are formulated for each of the following cases: (i) U convex, (ii) $f(x, U, t)$ a convex set, and (iii) U open.

Remark. The following classification of high order necessary conditions for optimality is possible. On the one hand, following the classical calculus of variations, conditions that use only the first variation of a functional may be called first order conditions; criteria based on the second variation may be called second order conditions, etc. On the other hand, one may follow optimal control theory, and base classification not upon classical variations of the form $\varepsilon\, \delta u(t)$, $t \in T$, but upon variations (of needle type) in which the parameter ε characterizes the length of the interval on which the variation is different from zero. With such an approach, the first order conditions are said to be those criteria which are obtained by considering the coefficient of the first power of ε in the expansion of

the increment of the cost functional. Similarly, criteria obtained by considering the coefficient of ε^2 are said to be second order conditions, etc. In this section, we have made use of precisely this classification.

JOINING OF EXTREMALS

9. Joining of singular and nonsingular controls. So far we have studied only singular controls and we have not at all touched upon the question of imbedding singular arcs into an entire optimal trajectory. But a singular control is, as a rule, a part of an optimal one. In this connection, the problem of an optimal joining of singular and nonsingular arcs of a trajectory arises. In other words, it is necessary to find properties which distinguish an optimal control from the remaining controls at the junction points of singular and nonsingular arcs. These problems, to some extent or other, were touched upon in [13], [14] and [63]–[65].

9.1. Let us consider a system with a one-dimensional control:

$$\dot{x} = f_0(x) + uf_1(x), \qquad x(t_0) = x_0,$$
$$|u(t)| \leqq \mathscr{L}, \quad \mathscr{L} > 0, \quad t \in T = [t_0, t_1].$$

Suppose that the functional (3) is to be minimized on the trajectories of this system which are generated by piecewise-continuous and piecewise-smooth (from the right) controls.

The Hamiltonian function for the given problem has the form

$$H(x, \psi, u) = \psi' f_0(x) + u\psi' f_1(x).$$

If $u(t)$, $t \in T$, is an admissible control which is singular on an arc $\sigma \subset T$, i.e., if $\partial H/\partial u = 0$, $t \in \sigma$, then, for this control to be optimal, it is necessary that the following conditions be satisfied:

(i) $u(t) = \mathscr{L} \operatorname{sgn} \psi'(t) f_1(x(t)), t \in T - \sigma$

(the maximum principle (6)),

(ii) on the arc σ,

$$(-1)^k \frac{\partial}{\partial u}\left(\frac{d^{2k}}{dt^{2k}} \frac{\partial H}{\partial u}\right) \leqq 0, \qquad k = 1, 2, \cdots \tag{95}$$

(see (47)).

9.2. Let $\theta \in T$ be a junction point of a singular and a nonsingular arc of an optimal control; i.e., let the control $u(t)$ be singular (in the classical sense) in a left-hand side neighborhood of the point θ, and nonsingular in a right-hand side neighborhood. We assume that the junction has the following properties:

(a) $u(t)$ is discontinuous at $t = \theta$,

(b) the smallest value of the index k for which the left-hand side of (95) is negative at $t = \theta$ equals q.

$$(-1)^q \frac{\partial}{\partial u}\left(\frac{d^{2q}}{dt^{2q}} \frac{\partial H}{\partial u}\right)\Bigg|_{t=\theta} < 0. \tag{96}$$

(c) $u(\theta) = u(\theta + 0) = \mathscr{L}$.

It follows from property (c) that there exists a $\delta > 0$ such that

$$\partial H/\partial u > 0, \qquad t \in (\theta, \theta + \delta).$$

In other words, the first nonzero derivative $(d^k/dt^k)(\partial H/\partial u)$ is positive at $t = \theta$. Taking this into account, property (b) implies

$$\begin{aligned} &\frac{d^k}{dt^k}\frac{\partial H}{\partial u}\bigg|_{t=\theta} = 0, && k = 0, 1, \cdots, 2q-1, \\ &\frac{d^{2q}}{dt^{2q}}\frac{\partial H}{\partial u} = [a(x,\psi) + ub(x,\psi)]_{t=\theta} > 0. \end{aligned} \tag{97}$$

On the other hand, by virtue of the singularity of $u(t)$, we have

$$\frac{d^{2q}}{dt^{2q}}\frac{\partial H}{\partial u} = [a(x,\psi) + ub(x,\psi)]_{t=\theta-0} = 0. \tag{98}$$

From (97) and (98), as a consequence of the continuity of the functions $x(t)$ and $\psi(t)$ and of property (a), we obtain

$$\frac{\partial}{\partial u}\left(\frac{d^{2q}}{dt^{2q}}\frac{\partial H}{\partial u}\right)\bigg|_{t=\theta} > 0. \tag{99}$$

If q is even, then the result (99) contradicts the assumption (96).

Thus, for the described junction to be optimal, it is necessary that q in (96) be odd [13], [14] and [65].

It can be shown by similar considerations [65] that this result remains in force also for all of the other kinds of junctions (singular–nonsingular, or nonsingular–singular controls), if the control is discontinuous at the junction point.

9.3. Let the following property be satisfied at a junction point $t = \theta$:

(a) $t = \theta$ is a corner point of a continuous control $u(t)$, and properties (b) and (c) of § 9.2 hold.

Then, on the one hand,

$$\frac{d^{2q+1}}{dt^{2q+1}}\frac{\partial H}{\partial u} = \left[\frac{d}{dt}a(x,\psi) + u\frac{d}{dt}b(x,\psi) + \dot{u}b(x,\psi)\right]_{t=0} > 0.$$

On the other hand,

$$\frac{d^{2q+1}}{dt^{2q+1}}\frac{\partial H}{\partial u} = \left[\frac{d}{dt}a(x,\psi) + u\frac{d}{dt}b(x,\psi) + \dot{u}b(x,\psi)\right]_{t=\theta-0} = 0.$$

By virtue of the continuity of the functions $x(t)$, $\psi(t)$, and $u(t)$ at $t = \theta$ we conclude that

$$\frac{\partial}{\partial u}\left(\frac{d^{2q}}{dt^{2q}}\frac{\partial H}{\partial u}\right)\bigg|_{t=\theta} < 0.$$

If q is odd, then the just-obtained inequality contradicts (96). The same contradiction is obtained also for other types of junctions, provided that, at the junction time, the control is continuous but has a corner point [14].

Conclusion: For the junction considered in this section to be optimal, it is necessary that q in (96) be even.

Remark. At a junction point $t = \theta$, let a control $u(t)$ be such that

$$u(\theta - 0) = u(\theta), \qquad \dot{u}(\theta - 0) = \dot{u}(\theta), \cdots, u^{(p-1)}(\theta - 0) = u^{(p-1)}(\theta), \tag{100}$$
$$u^{(p)}(\theta - 0) \neq u^{(p)}(\theta).$$

Moreover, let the singular control satisfy (96). Completely analogously to what was previously done, we may arrive at the following conclusion: If the number p in (100) is even (respectively, odd), then, for the control $u(t)$ to be optimal at the junction point $t = \theta$, it is necessary that q in (96) be odd (respectively, even).

9.4. In §9.2 and §9.3, we obtained conditions which were imposed on the order of the singularity of an optimal control on a singular arc, when there was a junction point with a nonsingular arc. Another method of investigating optimality conditions at junctions was proposed in [57], where an increment of the functional from control variations given at the junction point was considered. The conditions for optimality of junctions of singular and nonsingular controls consist of the following:

Let us consider the system

$$\dot{x} = f(x, u, t), \qquad u(t) \in U, \quad t \in T,$$

where U is an open set in the real line.

A singular (in the classical sense) control is characterized in this case by the identities

$$\partial H/\partial u \equiv 0, \qquad \partial^2 H/\partial u^2 \equiv 0.$$

(i) Let $\theta \in T$ be a junction point of a singular and a nonsingular arc of a control. For a junction to be optimal, it is necessary that the following conditions be satisfied:

(a) $$\partial^2 H/\partial u^2|_{t=\theta+0} \leqq 0$$

(the left-hand side is evaluated from the nonsingular control side);

(b) if

$$\partial^2 H/\partial u^2|_{t=\theta+0} = 0$$

(this always holds for systems with the control entering quadratically), then

$$\left[2\left(\frac{\partial^2 H'}{\partial u \partial x}\frac{\partial f}{\partial u} + \frac{\partial f'}{\partial u}\Psi\frac{\partial f}{\partial u}\right) + \frac{d}{dt}\frac{\partial^2 H}{\partial u^2}\right]_{t=\theta+0} \leqq 0;$$

(c) if (a) and (b) are ineffective, then

$$\left[\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right) + 3\frac{d}{dt}\left(\frac{\partial^2 H'}{\partial u \partial x}\frac{\partial f}{\partial u} + \frac{\partial f'}{\partial u}\Psi\frac{\partial f}{\partial u}\right)\right]_{t=\theta+0} \leqq 0. \tag{101}$$

(ii) Let us consider the case of a junction of a nonsingular and a singular arc. For a control $u(t)$ to be optimal at $t = \theta$, it is necessary that the following conditions be satisfied:

(a) $$\partial^2 H/\partial u^2|_{t=\theta-0} \leqq 0;$$

(b) if (a) is ineffective, then

$$\left[2\left(\frac{\partial^2 H'}{\partial u \partial x}\frac{\partial f}{\partial u}+\frac{\partial f'}{\partial u}\Psi\frac{\partial f}{\partial u}\right)-\frac{d}{dt}\frac{\partial^2 H}{\partial u^2}\right]_{t=\theta-0}\leqq 0;$$

(c) if the left-hand sides of the inequalities in (a) and (b) vanish, then

$$\left[\frac{\partial}{\partial u}\left(\frac{d^2}{dt^2}\frac{\partial H}{\partial u}\right)-3\frac{d}{dt}\left(\frac{\partial^2 H'}{\partial u \partial x}\frac{\partial f}{\partial u}+\frac{\partial f'}{\partial u}\Psi\frac{\partial f}{\partial u}\right)\right]_{t=\theta-0}\leqq 0. \tag{102}$$

One does not need to know the singular control in order to apply the conditions presented in this section. This may turn out to be useful in problems in which it is difficult to evaluate a singular control.

The conditions for an optimal junction obtained in § 9.2 and § 9.3 were used in [63] to investigate the possibility of including intermediate-thrust arcs in an optimal trajectory of a rocket flight.

Remark. Necessary conditions for optimal junctions were obtained in [64], expressed in terms of the continuity of certain functions.

10. Other types of junctions.

10.1. Utilizing the results of § 7, it is not difficult to obtain conditions for optimal junctions of singular controls of different orders of singularity. For example, let $\theta \in T$ be a junction point of singular controls which are such that, for one of them, $N(t) \equiv 0$, while for the other one, $N(t) \neq 0$. If the first control is defined for $t < \theta$, and the second one for $t > \theta$, then, for their junction to be optimal, it is necessary that inequality (101) hold. In the opposite case, if $N(t) \neq 0$ for $t < \theta$ and $N(t) \equiv 0$ for $t > \theta$, then inequality (102) is to be used. Similarly, with the aid of condition (82) of § 7, a criterion for an optimal junction of two singular controls whose orders of singularity differ more significantly can be formulated.

10.2. Let us consider the case of the junction of two nonsingular controls. We shall restrict ourselves to the case where the junction time $t = \theta$ is similar to a switching point of a nonsingular control. In other words, let the equation

$$\Delta_v H(x(\theta), \psi(\theta), u(\theta), \theta) = 0 \tag{103}$$

hold for certain $v \in U$ (U bounded).

The formula for the increment of a functional in a neighborhood of θ yields the following necessary condition for optimality of the switching times of a control:

$$\left[\left(\frac{\partial \Delta_v H}{\partial x}\right)'\Delta_v f+\Delta_v f'\Psi\Delta_v f\pm\frac{d}{dt}\Delta_v H\right]_{t=\theta\pm 0}\leqq 0 \tag{104}$$

for all $v \in U$ for which (103) holds.

Example 11.

$$\dot{x}_1 = u, \quad \dot{x}_2 = -x_1^2, \quad x_1(0) = x_2(0) = 0,$$
$$T = [0, 1], \quad |u| \leqq 1, \quad \phi(x) = x_2.$$

In this example, all of the controls (i) $u^*(t) \equiv 0$, and (ii) $u^p(t) = \pm \operatorname{sgn} \cos 3^p \pi t/2$, $p = 0, 1, 2, \cdots$, satisfy the maximum principle. The control $u^*(t)$ is singular and, as was shown in § 3, is not optimal. Each of the controls $u^p(t)$, $p \geqq 1$, has, on the interval $[0, 1]$, at least one switching point at which condition (103) holds. Applying condition (104) at such points shows that none of the controls $u^p(t)$, $p \geqq 1$, can be optimal. Thus, condition (104) singles out, from the infinity of controls which satisfy the maximum principle, only the two controls $u^0(t) = \pm 1$, which are, indeed, optimal for the problem under consideration.

NONSINGULAR CONTROLS

In the classical calculus of variations, there are no results on high order necessary conditions for optimality of singular extremals. But there is a series of very effective criteria for nonsingular extremals. First of all, there is the necessary condition of Jacobi, by virtue of which there must be no conjugate points between the endpoints of an extremal. Other conditions, expressed in terms of the second variation, are more complicated to verify. We shall not dwell on these results because, first of all, they are contained in any course on the calculus of variations, and, second of all, these results lie slightly to the side of the mainstream of this survey.

11. The case of an open control region.

11.1. In the theory of optimal processes, no constraints are imposed on the set U of values of the admissible controls, and the cases in which these sets are closed are considered to be typical. Therefore, as the fundamental definition of a singular control, one should consider Definition 1 of a control which is singular in the sense of the maximum principle. This does not contradict the frequently encountered fact that, in many concrete problems, optimal controls on singular arcs pass through interior points of the set U. Our selection of the basic type of singular control was brought forth only by a tendency to emphasize those methods of investigation which enable us to take into account different possible situations, such as when U does not contain any interior points, or when the control lies on the boundary of U, etc.

In studying high order necessary conditions for controls which are not singular in the sense of the maximum principle, one encounters serious difficulties. This is explained by the fact that the maximum principle itself is a very strong first order condition. In this section, we shall make use of some peculiarities of optimal controls which are not singular in the sense of the maximum principle, and of the structure of U near a point corresponding to an optimal control.

11.2. As we noted in § 1, a control which is singular in the sense of Definition 1 is also singular in the sense of Definition 2. However, it may turn out that a control which is singular in the classical sense is not singular in the sense of the maximum principle. In this case, to investigate controls which are nonsingular in the sense of Definition 1, one may use the fact that high order conditions for controls which are singular in the classical sense are stronger than the maximum principle.

For example, if the maximum principle is satisfied (without singularity) by a control $u(t)$, but this control is singular in the classical sense, then, to verify the optimality of $u(t)$, one can use all of the necessary conditions which were obtained

for controls which are singular in the sense of Definition 2. Example 1 shows that such a situation may arise and that the method will turn out to be effective. In this example, the control $u(t) \equiv 0$ is nonsingular in the sense of Definition 1. The Hamiltonian of the system attains its strict absolute maximum at $u = 0$, i.e., the maximum principle cannot exclude this control as a possible candidate for an optimal control. On the other hand, the control $u(t) \equiv 0$ is singular in the classical sense. Calling on condition (40) we can convince ourselves that this control is not optimal.

12. Problems with closed control sets. Let us consider second order conditions which are based on peculiarities of the structure of the set U in a neighborhood of an optimal control.

12.1. Let us consider the formula for the second order increment (92). Let $\tilde{u}(t)$ be generated by a needle variation:

$$\tilde{u} = \begin{cases} u(t), & t \notin [\theta, \theta + \varepsilon), \\ v(\varepsilon), & t \in [\theta, \theta + \varepsilon), \end{cases}$$

where the vector $v(\varepsilon) \in U$ is chosen such that

$$\Delta_{v(\varepsilon)} f(x, u(t), t) = \varepsilon a_1(x, t) + \varepsilon^2 a_2(x, t) + \varepsilon^3 a_3(x, t).$$

The parameters $a_i = a_i(x, t)$, $i = 1, 2, 3$, will be said to be admissible at the point x, θ if

$$f(x, v(\varepsilon_i), \theta) \in f(x, U, \theta)$$

for all ε_i, $i \geqq N_0$ ($\varepsilon_i > 0$ and $\varepsilon_i \to 0$ as $i \to \infty$).

Expanding the increment of the cost functional in powers of ε, we obtain the following assertions [62].

If, for some $t = \theta \neq t_1$, and for an admissible parameter a_1, the equation

$$\psi'(\theta) a_1(x(\theta), \theta) = 0 \tag{105}$$

holds, then, for $u(t)$ to be optimal, it is necessary that the inequality

$$\frac{d}{dt}[\psi'(t) a_1(x(t), t)]_{t=\theta+0} + 2\psi'(\theta) a_2(x(\theta), \theta) \leqq 0 \tag{106}$$

hold for all admissible a_1 and a_2.

If, for some admissible a_1 and a_2, both conditions (105) and (106) are satisfied at $t = \theta \neq t_1$ as equalities, then a necessary condition for $u(t)$ to be optimal is that the inequality

$$\begin{aligned} &\frac{d^2}{dt^2}[\psi'(t) a_1(x(t), t)]_{t=\theta+\varepsilon} + 3\frac{d}{dt}[\psi'(t) a_2(x(t), t)]_{t=\theta+0} \\ &+ 6\psi'(\theta) a_3(x(\theta), \theta) + 3\frac{\partial[\psi(\theta) a_1(x(\theta), \theta)]'}{\partial x} a_1(x(\theta), \theta) \\ &+ 3a_1'(x(\theta), \theta)\Psi(\theta) a_1(x(\theta), \theta) \leqq 0 \end{aligned} \tag{107}$$

hold for all admissible a_1, a_2, and a_3.

Conditions (106) and (107) become simplified [62] if the set of admissible velocities $f(x, U, \theta)$ in system (1) is convex.

12.2. In the preceding, the structure of U was taken into account indirectly, through the set $f(x, U, t)$. In what follows, conditions will be presented which take into account the structure of U directly. Let the function $f(x, u, t)$ be twice differentiable with respect to u. Then the formula for the increment (92) can be simplified by explicitly separating out the control variations. Considering these variations to be of needle type, we set

$$\Delta u(t) \equiv v(\varepsilon) - u(t) = \begin{cases} 0, & t \notin [\theta, \theta + \varepsilon), \\ \varepsilon b_1(t) + \varepsilon^2 b_2(t) + \varepsilon^3 b_3(t), & t \in [\theta, \theta + \varepsilon). \end{cases}$$

The parameters $b_i = b_i(t)$, $i = 1, 2, 3$, will be said to be admissible if $v(\varepsilon_i) \in U$ for all ε_i, $i \geqq N$.

For an admissible b_1, let

$$\frac{\partial[\psi'(\theta) f(x(\theta), u(\theta), \theta)]'}{\partial u} b_1(\theta) = 0. \tag{108}$$

Then, for this $b_1(\theta)$ and for all admissible $b_2(\theta)$ with an optimal $u(\theta)$, the inequality

$$\frac{d}{dt}\left[\frac{\partial(\psi' f)'}{\partial u} b_1\right]_{t=\theta+0} + 2\frac{\partial(\psi' f)'}{\partial u} b_2(\theta) + b_1'(\theta)\frac{\partial^2(\psi' f)}{\partial u^2} b_1(\theta) \leqq 0 \tag{109}$$

holds.

But if, for admissible $b_1(\theta)$ and $b_2(\theta)$, conditions (108) and (109) hold as equalities, then, for all these b_1 and b_2 and all admissible $b_3(\theta)$ with an optimal $u(\theta)$, the following inequality is satisfied:

$$\frac{d^2}{dt^2}\left[\frac{\partial(\psi' f)'}{\partial u} b_1\right]_{t=\theta+0} + 3\frac{d}{dt}\left[\frac{\partial(\psi' f)'}{\partial u} b_2\right]_{t=\theta+0} + 6\frac{\partial(\psi' f)'}{\partial u} b_3(\theta)$$

$$+\frac{3}{2}\frac{d}{dt}\left[b_1'\frac{\partial^2(\psi' f)}{\partial u^2} b_1\right]_{t=\theta+0} + 6b_1'\frac{\partial^2(\psi' f)}{\partial u^2} b_2 + 3b_1'\frac{\partial^2(\psi' f)}{\partial u \partial x}\frac{\partial f}{\partial u} b_1$$

$$+ 3b_1'\frac{\partial f'}{\partial u}\Psi\frac{\partial f}{\partial u} b_1 \leqq 0.$$

The convexity assumption on U allows one [62] to obtain additional, simpler criteria.

12.3. We shall illustrate the results of § 12.1 and § 12.2 with an example.

Example 12.

$$\dot{x}_1 = u_1, \quad \dot{x}_2 = -x_1^2 + u_2, \quad x_1(0) = x_2(0) = 0, \quad T = [0, 1],$$

$$U = \{u_1, u_2 : |u_1| \leqq 1, |u_2| \leqq \tfrac{1}{2}, u_2 \geqq \tfrac{1}{2}|u_1|^3\},$$

$$J(u) = x_2(1).$$

The control $u_1(t) = u_2(t) \equiv 0$ satisfies the maximum principle, and, moreover, the Hamiltonian attains its maximum at the single point $u_1 = u_2 = 0$. The admissible parameters a_i (b_i) are the following:

$$a_1 = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix}, \quad a_2 = \begin{Bmatrix} 0 \\ \alpha \end{Bmatrix}, \quad a_3 = \begin{Bmatrix} 0 \\ \beta \end{Bmatrix}, \quad \alpha \geqq 0, \quad \beta \geqq \tfrac{1}{2}.$$

For an admissible a_1, (105) holds. For $\alpha = 0$, (106) also holds. Inequality (107) has the form

$$-6\beta + 6(1 - \theta) \leqq 0,$$

and, obviously, cannot be satisfied for all θ, $0 \leqq \theta \leqq 1$, $\beta \geqq \frac{1}{2}$.

Thus, the control under consideration is not optimal.

Remark. Slightly to the side with respect to the methods of derivation lie the high order optimality conditions presented in [66].

CONCLUSION

To summarize this survey of the present state of the problem of high order necessary conditions for optimality, we note the following circumstances.

1. The problem owes its origin to a series of important applied problems. Although the majority of the known problems of this type are associated with space navigation, one must assume that in other domains the importance of the new problems will not be any the less.

2. The methods of solution for the problems are only being conceived. Here one can use the results of the classical calculus of variations to an even lesser extent than was the case when investigating first order conditions. A high order "maximum principle" still awaits discovery.

The basic results which were obtained in the theory of high order necessary conditions for optimality pertain to optimization problems with a free right-hand endpoint. Although these results are quite sufficient in order to construct many computational algorithms, the theory of problems with fixed and partially fixed endpoints is of interest in itself. A general scheme for taking into account additional constraints was described in [4]. Studies of boundary conditions for high order necessary conditions were presented also in [14], [34] and [37], but this problem is in need of additional investigation.

3. As in the case of first order conditions, the new results are bound to influence the practical aspects of computing optimal controls. Various computational schemes which are based on first order conditions (which we shall call first order algorithms) often present some "peculiarities" when they are realized on electronic computers. A common feature of such schemes is the fact that their convergence becomes worse as one approaches an optimum. We may hope that algorithms based on high order conditions will have lesser drawbacks. Undoubtedly, the situation here is substantially more complicated than in the finite-dimensional case [67]–[71], but, doubtless, attempts at considering high order algorithms will prove useful.

REFERENCES

[1] V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND L. S. PONTRYAGIN, *On the theory of optimal processes*, Dokl. Akad. Nauk SSSR, 110 (1956), pp. 7–10 (in Russian).

[2] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[3] G. A. BLISS, *Lectures on the Calculus of Variations*, Univ. of Chicago, Chicago, 1946.

[4] A. YA. DUBOVITSKII AND A. A. MILYUTIN, *Extremum problems in the presence of restrictions*, USSR Comput. Math. and Math. Phys., 5 (1965), pp. 1–80.

[5] R. V. GAMKRELIDZE, *On some extremal problems in the theory of optimal control*, this Journal, 3 (1965), pp. 106–128.

[6] L. 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.

[7] A. I. LUR'E AND A. V. TROITSKII, *The problem of Mayer and Bolza, and optimal control processes*, Proc. Fourth Math. Congress, vol. II, Nauka, Moscow, 1964, pp. 203–227 (in Russian).

[8] R. E. KALMAN, *The theory of optimal control and the calculus of variations*, Mathematical Optimization Techniques, R. Bellman, ed., University of California, Berkeley, 1963, pp. 309–331.

[9] R. V. GAMKRELIDZE AND G. L. KHARATISHVILI, *Extremal problems in linear topological spaces*, Math. USSR-Izv., 3 (1969), pp. 737–794.

[10] L. I. ROZENOER, *L. S. Pontryagin's maximum principle in the theory of optimum systems. I*, Automat. Remote Control, 20 (1959), pp. 1288–1302.

[11] ———, *L. S. Pontryagin's maximum principle in optimal system theory. II*, Ibid., 20 (1959), pp. 1405–1421.

[12] E. J. MCSHANE, *On multipliers for Lagrange problems*, Amer. J. Math., 61 (1939), pp. 809–819.

[13] C. D. JOHNSON, *Singular solutions in problems of optimal control*, Advances in Control Systems. Theory and Applications, vol. 2, C. T. Leondes, ed., Academic Press, New York, 1965, pp. 209–267.

[14] H. J. KELLEY, R. E. KOPP AND H. G. MOYER, *Singular extremals*, Topics in Optimization, Academic Press, New York, 1967, pp. 63–101.

[15] R. V. GAMKRELIDZE, *Optimal sliding states*, Soviet Math. Dokl., 3 (1962), pp. 559–561.

[16] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–145.

[17] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[18] R. GABASOV, *On the theory of necessary conditions for optimality of singular controls*, Dokl. Akad. Nauk SSSR, 183 (1968), pp. 300–302 (in Russian).

[19] J. C. DUNN, *On the classification of singular and nonsingular extremals for the Pontryagin maximum principle*, J. Math. Anal. Appl., 17 (1967), pp. 1–36.

[20] YU. I. PARAEV, *On singular control in optimal processes that are linear with respect to the control inputs*, Automat. Remote Control, 23 (1962), pp. 1127–1134.

[21] C. D. JOHNSON AND J. E. GIBSON, *Singular solutions in problems of optimal control*, IEEE Trans. Automat. Control, AC-8 (1963), pp. 4–15.

[22] M. ATHANS AND P. L. FALB, *Optimal Control*, McGraw-Hill, New York, 1966.

[23] A. I. LUR'E, *Analytical Mechanics*, Fizmatgiz, Moscow, 1961 (in Russian).

[24] A. M. LETOV, *Flight Dynamics and Control*, Nauka, Moscow, 1969 (in Russian).

[25] W. M. WONHAM AND C. D. JOHNSON, *Optimal bang-bang control with quadratic index of performance*, Trans. ASME.Ser. D. J. Basic Engrg., 1 (1964), pp. 107–115.

[26] V. I. BUYAKAS, *Optimal control by (sic) systems with variable structure*, Automat. Remote Control, 27 (1966), pp. 579–589.

[27] ———, *Optimal control of systems with variable structure having constantly acting disturbances*, Ibid., 27 (1966), pp. 1226–1235.

[28] B. S. GOH, *Optimal singular control for multi-input linear systems*, J. Math. Anal. Appl., 20 (1967), pp. 534–539.

[29] R. A. ROHRER AND M. SOBRAL, *Optimal singular solutions for linear multi-input systems*, Trans. ASME Ser. D. J. Basic Engrg., 88 (1966), pp. 323–328.

[30] H. HERMES AND G. W. HAYNES, *On the nonlinear control problem with control appearing linearly*, this Journal, 1 (1963), pp. 85–108.

[31] A. A. BOLONKIN, *A method for solving optimal problems*, Complex Control Systems, Naukova Dumka, Kiev, 1965, pp. 65–69 (in Russian).
[32] A. M. CHECHEL'NITSKII, *Dynamics of singular optimal motions*, Engrg. Cybernetics, 2 (1969), pp. 1–11.
[33] ———, *Optimal motions with many singular controls and reduced systems*, Ibid., 4 (1969), pp. 32–36.
[34] H. J. KELLEY, *A second variation test for singular extremals*, AIAA J., 2 (1964), pp. 1380–1382.
[35] R. GABASOV, *One problem in the theory of optimal processes*, Automat. Remote Control, 28 (1967), pp. 1085–1094.
[36] C. T. STREIBEL AND J. V. BREAKWELL, *Minimum effort control in interplanetary guidance*, IAS Preprint, 1963, pp. 63–80.
[37] R. E. KOPP AND H. G. MOYER, *Necessary conditions for singular extremals*, AIAA J., 3 (1965), pp. 1439–1444.
[38] D. F. LAWDEN, *Optimal Trajectories for Space Navigation*, Butterworths, London, 1963.
[39] B. S. GOH, *The second variation for the singular Bolza problem*, this Journal, 4 (1966), pp. 309–325.
[40] ———, *Necessary conditions for singular extremals involving multiple control variables*, this Journal, 4 (1966), pp. 716–731.
[41] I. B. VAPNYARSKII, *An existence theorem for optimal control in the Boltz (sic) problem, some of its applications and the necessary conditions for the optimality of moving (sic) and singular systems*, USSR Comput. Math. and Math. Phys., 7 (1967), pp. 22–54.
[42] A. A. BOLONKIN, *Special extremals in optimal control problems*, Engrg. Cybernetics, 2 (1969), pp. 170–182.
[43] J. V. BREAKWELL, *A doubly singular problem in optimal interplanetary guidance*, this Journal, 3 (1965), pp. 71–77.
[44] H. J. KELLEY, *A transformation approach to singular subarcs in optimal trajectory and control problems*, this Journal, 2 (1964), pp. 234–240.
[45] V. I. GURMAN, *Optimal processes of singular control*, Automat. Remote Control, 26 (1965), pp. 783–792.
[46] ———, *A method for the investigation of one class of optimal sliding states*, Ibid., 26 (1965), pp. 1159–1166.
[47] ———, *Method of multiple maxima and the conditions of relative optimality of degenerate regimes*, Ibid., 28 (1967), pp. 1845–1852.
[48] ———, *The method of multiple maxima and optimization problems for space maneuvers*, Proc. Second Readings of K. E. Tsiolkovskii (Kaluga, 1967), Moscow, 1968, pp. 39–51 (in Russian).
[49] F. G. TRICOMI, *Equazoni a Derivate Parziali*, Cremonese, Rome, 1957.
[50] V. I. GURMAN AND A. M. NIKULIN, *Problems of realizing an optimal cyclic regime for stopping the rotation of a satellite*, Proc. Third Readings of K. E. Tsiolkovskii (Kaluga, 1968), Mechanics of Space Flight Section, U.S.S.R. Academy of Sciences Publishing House, Moscow, 1969, pp. 45–62 (in Russian).
[51] A. M. NIKULIN, *The realization of a motion along a curve of zero-distance of sliding optimal regimes*, Izv. Vyssh. Uchebn. Zaved. Aviats. Tekh., 4 (1969), pp. 17–24 (in Russian).
[52] V. I. GURMAN, *On optimal trajectories of a jet aircraft in a central field*, Kosmicheskie Issledovaniya, 3 (1965), pp. 368–373 (In Russian).
[53] ———, *On optimal transitions between coplanar elliptical orbits in a central field*, Ibid., 4 (1966), pp. 26–39 (in Russian).
[54] ———, *On the problem of optimality of singular regimes of rocket motions in a central field*, Ibid., 4 (1966), pp. 499–509 (in Russian).
[55] ———, *The structure of optimal regimes of rocket motions in a homogeneous gravitational field*, Ibid., 4 (1966), pp. 815–822 (in Russian).
[56] R. GABASOV, F. M. KIRILLOVA AND V. A. SROCHKO, *On the theory of optimal singular controls*, Abstracts of the Fifth International Conference of Non-linear Oscillations, Kiev, 1969, pp. 59–60 (in Russian).
[57] R. GABASOV AND V. A. SROCHKO, *Investigation of singular controls with the aid of a variation bundle*, Differentsial'nye Uravnenija, 6 (1970), pp. 260–275 (in Russian).
[58] V. A. SROCHKO, *The connection between two necessary conditions for optimality of singular controls*, Ibid., 6 (1970), pp. 387–389 (in Russian).

[59] R. GABASOV AND F. M. KIRILLOVA, *Optimal control with singular regions (sic)*, Automat. Remote Control, 30 (1969), pp. 1554–1563.

[60] R. GABASOV, *Necessary conditions for optimality of singular control*, Engrg. Cybernetics, 5 (1968), pp. 28–36.

[61] ———, *On optimality of singular controls*, Differentsial'nye Uravnenija, 4 (1968), pp. 1000–1011 (in Russian).

[62] R. GABASOV AND F. M. KIRILLOVA, *On the theory of high-order necessary conditions for optimality*, Ibid., 6 (1970), pp. 665–676 (in Russian).

[63] H. M. ROBBINS, *Optimality of intermediate-thrust arcs of rocket trajectories*, AIAA J., 3 (1965), pp. 1094–1098.

[64] YU. A. KLIKH, O. F. MAKAROV AND V. A. PLOTNIKOV, *On joining singular and non-singular extremals*, Dopovidi Akad. Nauk Ukrain. RSR, 9 (1967), pp. 778–782 (in Ukrainian).

[65] ———, *On imbedding singular arcs into an optimal trajectory*, Mat. Fiz. Respublik. Mezhvedomstvennyi Sbornik, No. 4, Kiev, 1968, pp. 42–49 (in Russian).

[66] R. GABASOV AND F. M. KIRILLOVA, *A maximum principle for extremals of L. S. Pontryagin*, Differentsial'nye Uravnenija, 4 (1968), pp. 963–972 (in Russian).

[67] R. FLETCHER AND M. J. POWELL, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.

[68] A. I. KOIVUNIEMI AND W. E. HAMILTON, *An algorithm for the solution to an optimal singular control problem*, Proc. National Electronics Conference (Chicago, 23 (1967)), Chicago, 1967, pp. 20–94.

[69] V. F. DEMYANOV AND A. M. RUBINOV, *Approximate Methods in Optimization Problems*, American Elsevier, New York, 1970.

[70] G. M. OSTROVSKII, YU. M. VOLIN AND V. V. BORISOV, *Optimization of complex systems with the aid of a second-order method*, Seeking an Extremum, Tomsk, 1969, pp. 301–306 (in Russian).

[71] W. F. POWERS, *Techniques for improved convergence in neighboring optimum guidance*, AIAA J., 8 (1970), pp. 2235–2241.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

SUR L'EXISTENCE ET LE COMPORTEMENT DE CYCLES DE CERTAINES ÉQUATIONS DIFFÉRENTIELLES NON LINÉAIRES (SOLUTIONS PERIODIQUES DE SYSTEMES AUTONOMES)*

ROBERT FAURE†

Abstract. This paper deals with the properties of cycles and of their frequency ω for some nonlinear differential equations depending on a parameter λ (second order differential equation, coupling of two second order equations, and third order equation).

The use of inequalities from the theory of series and of integrals permits us to establish some norms of the solutions as a function of ω, to relate λ and ω, and to study the dependence of the norms on λ and ω.

The equations, once they have been put in the form of a transformation on a function space, allow us, through the use of the Leray–Schauder theorem, to prove, on the basis of results for small $|\lambda|$, the existence of cycles for all values of the parameter. We sketch what happens to the cycles as $|\lambda| \to \infty$.

Introduction. Le problème des cycles est un des problèmes importants de la Mécanique et de l'Electronique non linéaires.

Nous allons étudier dans ce qui suit certaines équations différentielles non linéaires.

Un des éléments de base de cette étude qui utilise le théorème de Leray–Schauder est la connaissance d'un théorème d'existence de "cycle" pour des équations différentielles linéaires perturbées non linéairement que nous schématisons par:

$$Lx = \lambda F(x, x'), \quad \lambda \in R,$$

où L est un opérateur différentiel linéaire. $F(x, x')$ est une fonction non linéaire de l'inconnue x et de sa dérivée première x', λ paramètre.

Une remarque importante doit être faite au sujet de l'application du théorème de Leray–Schauder à notre problème.

Pour une équation unique du deuxième ordre, le cycle représenté dans le plan des phases correspond à une courbe fermée $x(t)$, $x'(t)$ invariante dans une translation arbitraire h du temps. Il n'en est pas de même pour la solution $x(t)$; $x(t + h)$ est une autre solution qui admet une représentation de Fourier distincte de celle de $x(t)$.

Nous grouperons alors toutes les solutions dépendant de h en une unique dépendant de ω que nous désignerons sous le nom de "cycles" cette solution unique est associée à la pulsation unique ω qui jouera pendant la plus grande partie de l'étude le rôle de paramètre à la place de λ qui sera alors une inconnue. On procèdera de même pour tout système d'équations.

Nous utiliserons le théorème de Leray–Schauder tel qu'il est énoncé dans le mémoire original [6, p. 59, énoncé et remarque 20], [6, p. 63, Théorème et remarques H']. h le déphasage arbitraire et la pulsation sont alors deux paramètres; on voit

* Received by the editors April 21, 1970, and in final revised form January 6, 1971.

† Faculté des Sciences, Université de Dakar, Republique du Senegal. Now at 21 Avenue Gambetta, Castillon La Bataille, (33) Gironde, France.

aisément que h vérifie toutes les hypothèses nécessaires à l'application du théorème de Leray–Schauder. *On doit ajouter que les résultats obtenus sur les normes a priori des inconnues x et λ permettent l'application des théorèmes directs utilisant le degré topologique lorsque celui-ci est calculable.*

Dans la suite de l'exposé on désignera par H'_2 la deuxième partie des remarques H′ soit : on connait une valeur du paramètre k_0 pour laquelle l'équation admet un nombre impair de solutions au voisinage desquelles la transformation $X' = \mathscr{F}(x, k)$ est biunivoque.

Nous considérons des fonctions $u(t)$ périodiques de période T, la pulsation étant ω.

Nous posons comme norme de $u(t)$ désignée par $\|u\|$:

(i) si $u(t) \in H$ (espace de Hilbert),

$$\|u\|_H = \sqrt{\frac{1}{T}\int_0^T u^2\,dt};$$

(ii) si $u(t) \in B$ espaces des séries de Fourier absolument convergentes,

$$\|u\|_B = \sum_{n=-\infty}^{n=+\infty} |\alpha_n| \quad \text{avec} \quad u(t) = \sum_{n=-\infty}^{n=+\infty} \alpha_n e^{ni\omega t}.$$

1. Examinons tout d'abord les équations :

$$x'' + \omega_1^2 x = \lambda P(x'), \tag{E_1}$$

ω_1 est une constante, λ un paramètre positif.

On suppose de plus que $P(x')$ est un polynôme à coefficients constants de degré *impair* avec

$$P(x') = a_{2n+1}x'^{2n+1} + \cdots + a_1 x', \qquad a_1 \neq 0.$$

Nous allons supposer l'existence d'un cycle de période T (pulsation ω) et étudier les solutions correspondantes ; nous ramenons la période à 2π, en posant $\omega t = \tau$. Nous prenons alors pour nouvelle inconnue $u(\tau) = \omega x(\tau)$.

L'équation différentielle devient alors :

$$\omega^2 u''_{\tau^2} + \omega_1^2 u = \lambda\omega P(u'_\tau)$$

et *en reprenant les notations initiales x et t au lieu de u et τ, on a l'équation* :

$$\omega^2 x'' + \omega_1^2 x = \lambda\omega P(x'). \tag{E'_1}$$

Etudions (E'_1). Multiplions (E'_1) par x' et intégrons entre 0 et 2π.

On a :

$$\int_0^{2\pi} x' P(x')\,dt = 0 \tag{1.1}$$

qui fournit

$$\sum_{i=1}^{i=2n+1} a_i \int_0^{2\pi} x'^{i+1}\,dt = 0, \tag{1.2}$$

d'où l'inégalité :

$$|a_{2n+1}| \int_0^{2\pi} x'^{2n+2}\, dt \leqq \sum_{i \neq 2n+1} |a_i| \int_0^{2\pi} |x'^{i+1}|\, dt. \tag{1.3}$$

On déduit de (1.3) en posant $l^{2n+2} = (1/2\pi)\int_0^{2\pi} x'^{2n+2}\, dt$ et en utilisant l'inégalité de Hölder que:

$$|a_{2n+1}| l^{2n+2} \leqq \sum_{i \neq 2n+1} |a_i| l^{i+1}, \qquad i \neq 2n+1, \tag{1.4}$$

donc l est borné supérieurement par un nombre K_1 et a fortiori:

$$\|x'\|_H \leqq K_1.$$

Ceci est vrai également pour la fonction $x'(t)$ de période naturelle T solution de (E_1).

On en tire en outre, deux conclusions importantes :

$$\begin{aligned} &\text{(i)} \quad \frac{1}{2\pi}\int_0^{2\pi} |P(x')|\, dt \leqq \sum_{i=1}^{i=2p+1} |a_i| l^i = A \text{ const.}, \\ &\text{(ii)} \quad \frac{1}{2\pi}\int_0^{2\pi} |P(x')|^k\, dt \leqq B \text{ const.}, \quad k = \frac{2p+2}{2p+1}. \end{aligned} \tag{1.5}$$

Avec $a_1 \neq 0$, on remarque que $\int_0^{2\pi} P(x')x'\, dt$ étant nulle, il y a au moins un changement de signe sous le signe somme, $P(x')$ s'annule donc au moins une fois en dehors de zéro ; soit α une racine, on pose $|\alpha| = K_2$, d'où:

$$\begin{gathered} \|x'\|_B > K_2, \quad \int_0^{2\pi} |x''|\, dt \geqq K_2, \\ |x''|_H \geqq K_2/\sqrt{2\pi}. \end{gathered} \tag{1.6}$$

En multipliant (E_1') par x'' et intégrant par partie :

$$\omega^2 \int_0^{2\pi} x''^2\, dt - \omega_1^2 \int_0^{2\pi} x'^2\, dt = 0. \tag{1.7}$$

Vu (1.5) on tire de l'égalité précédente :

$$\omega \|x''\|_H \leqq \omega_1 K_1, \tag{1.8}$$

mais également :

$$\omega^2 \int_0^{2\pi} x''^2\, dt = \omega_1^2 \int_0^{2\pi} x'^2\, dt \leqq \omega_1^2 \int_0^{2\pi} x''^2\, dt,$$

d'où:

$$\omega^2 \leqq \omega_1^2. \tag{1.9}$$

On déduit de (1.8) la période de $x(t)$ étant 2π :

$$\|x'\|_B \leqq K_3/\omega ; \tag{1.10}$$

et par suite vu (1.6):

$$\|x'\|_H > K_4\omega. \tag{1.11}$$

K_1, K_2, K_3, K_4 sont des constantes positives.

1.1. Lien entre ω, λ et la solution x'. Etudions les différents cas limites:

(i) Nous supposons acquis les résultats de Poincaré et Haag [4] si $\lambda \to 0$ la pulsation du cycle tend alors vers ω_1 pulsation maximum, ce qui conduit ici à écrire:

$$\omega = \omega_1 - c_1\lambda^2 + \varepsilon\lambda^2, \qquad c_1 \geqq 0.$$

(ii) Réciproquement supposons $\omega \to \omega_1$.

Considérons maintenant avec $x_r = x - (1/2\pi)\int_0^{2\pi} x\,dt$ (E_1') sous la forme:

$$\omega^2 x'' + \omega_1^2 x_r = \lambda\omega\left(P(x') - \frac{1}{2\pi}\int_0^{2\pi} P(x')\,dt\right) = \lambda\omega Q(x'). \tag{1.12}$$

Calculons

$$\|\omega^2 x'' + \omega_1^2 x_r\|_H = \sqrt{\frac{1}{2\pi}\int_0^{2\pi} (\omega^2 x'' + \omega_1^2 x_r)^2\,dt}$$

égal vu (1.7) à:

$$\omega_1 \cdot \sqrt{\frac{1}{2\pi}\int_0^{2\pi} (\omega_1^2 x_r^2 - \omega^2 x'^2)\,dt} < \omega_1\sqrt{(\omega_1^2 - \omega^2)}\,\|x'\|_H,$$

d'où $\|\omega^2 x'' + \omega_1^2 x_r\|_H$ tend vers zéro. Par suite, $|\lambda|\omega\|Q(x')\|_H$ qui lui est égal tend vers zéro.

Montrons que l'ensemble des $\|Q(x')\|_H$ est borné inférieurement par un nombre M positif quand $\omega > \omega_0 > 0$, a fortiori si ω tend vers ω_1.

On a:

$$|x'(t_2) - x'(t_1)| = \left|\int_{t_1}^{t_2} x''\,dt\right| < \sqrt{|t_2 - t_1|\left|\int_{t_1}^{t_2} x''^2\,dt\right|}.$$

Par suite:

$$|x'(t_2) - x'(t_1)| \leqq \sqrt{2\pi}\frac{\omega_1 \cdot K_1}{\omega_0}|t_2 - t_1|^{1/2}.$$

Il en résulte deux conclusions:

(i) Let $x'(t)$ forment un ensemble de functions également continues.

(ii) Les $x'(t_1)$ pouvant s'annuler, on a en faisant $|t_2 - t_1| \leqq 2\pi$, $|x'(t_2)| \leqq 2\pi\omega_1 K_1/\omega_0$ donc bornés dans leur ensemble.

Supposons que $M = 0$ et ne soit pas atteinte, il existe alors une suite (ω_i, x_i') avec $\omega_i \to \omega_l$, alors que $\|Q(x')\|_H$ tend vers zéro.

D'après le théorème d'Arzela–Ascoli on peut en extraire une sous-suite x_i' tendant uniformément vers une limite $L(t)$ continue qui doit satisfaire la condition suivante:

$$\|Q(L(t))\|_H = M = 0.$$

Il en résulte que $P(L(t))$ a une partie oscillante nulle, $P(x'(t))$ tend donc uniformément vers une constante $K = P(L(t))$, $L(t)$ est constant et doit satisfaire ensus aux deux conditions suivantes: s'annuler et avoir $\max |L(t)| \geqq K_2$.

Il y a contradiction M ne peut être nul.

Il en est de même si M est effectivement atteint. Par suite dans les deux cas λ ne peut que tendre vers zéro si $\omega \to \omega_1$.

Montrons que lorsque λ tend vers $+\infty$, ω n'est pas borné inférieurement. *Supposons en effet que ω soit borné inférieurement par ω_0.*

Supposons qu'il en soit ainsi: Il résulte de l'étude faite au (ii) que $\|Q(x')\|_H$ est borné inférieurement par un nombre positif M.

On a toujours avec les notations du (ii):

$$\omega^2 x'' + \omega_1^2 x_r = \lambda\omega Q'(x)$$

or $\|\omega^2 x'' + \omega_1^2 x_r\|_H$, vu (1.5) et (1.8), est borné supérieurement par $2K_1\omega_1^2$; par suite on a:

$$2K_1\omega_1^2 \geqq \lambda\omega_0 M,$$

ou λ augmente indéfiniment, cela est impossible.

Il en résulte que ω ne peut être borné inférieurement lorsque λ augmente indéfiniment.

1.2. Démonstration de l'existence de solutions (C) quelque soit ω. La donnée de x' entrainant celle de x:

(i) On écrit l'équation différentielle sous la forme $X' = x'$; $(X', \Lambda) = S(x', \lambda)$ en adjoignant λ à x' comme inconnue, ω étant alors le paramètre de Leray–Schauder avec $0 < \alpha \leqq \omega \leqq \beta < \omega_1$, α, β *constantes.* S est définie de la manière suivante:

$$S = (1, T) = \begin{cases} \text{(a) } \Lambda = \lambda, \\ \text{(b) } X' = T(\lambda, x') = T(\lambda\omega P(x') + K\omega x'). \end{cases}$$

K est une constante réelle arbitraire choisie une fois pour toute. X', x' appartiennent tous deux à l'espace (B) des séries de période 2π, $K \neq 0$: $X', x' \in (B)$. T est la traduction dans (B) de l'équation différentielle:

$$\omega^2 X'' + K\omega X' + \omega_1^2 X = \lambda\omega P(x') + K\omega x' = \omega\Phi(x')$$

si

$$x' = \sum_{n=-\infty}^{n=+\infty} U_n e^{nit}, \qquad n \neq 0,$$

avec

$$\Phi(x') = \sum_{n=-\infty}^{n=+\infty} V_n e^{nit}.$$

On a alors:

$$X' = \sum_{n=-\infty}^{n=+\infty} t_n V_n e^{nit}$$

avec

$$t_n = \frac{ni}{\omega_1^2 - n^2 \cdot \omega^2 + Kn\omega i}.$$

L'intervention de K permet la définition de tous les t_n ainsi donc dans $(\mathscr{E})$ d'une transformation T pour tous les λ, ω, x' avec $\alpha \leqq \omega \leqq \beta$, $\mathscr{E} = R \times B$, $\lambda \in R$, $x' \in B$.

(ii) Vérifions que les conditions d'application du théorème [6, p. 63] sont vérifiées:

(a) Elles portent sur la continuité complète (continuité par rapport à l'argument et compacité) et sur la continuité uniforme par rapport au paramètre ω.

Bornons nous à vérifier la compacité ainsi que l'existence d'un domaine Ω borné pour les solutions.

Montrons que pour chaque valeur ω de $[I]_\alpha^\beta$ la transformation S transforme tout ensemble borné infini (F) de (λ, x') de $R \times B$ en un ensemble compact.

Ici $\lambda > 0$, la norme dans $(R \times B)$ est $|\lambda| + \|x'\|_B$. On a $|\lambda| + \|x'\|_B \leqq M$, où M est une constante.

Considérons tout d'abord une suite infinie $x'_i \in B$, on a $\|x'_i\|_B \leqq M$, on peut donc en extraire une sous-suite $x'_{i'} \in (B)$ qui converge faiblement vers un vecteur L_1.

Soit maintenant la suite $P(x'_i)$, cette suite infinie est bornée, ceci, vu l'inégalité $\|P(x'_{i'})\|_B \leqq Q(M)$ où $Q(M)$ est le polynome $Q(M) = \sum_{i=1}^{i=2n+1} |a_i| M^i$.

On peut donc en extraire une sous-suite infinie $P(x'_{i''})$ telle que $P(x'_{i''})$ converge elle aussi faiblement vers un vecteur L_2.

Pour λ le résultat est évident, il existe donc une suite $\lambda_{i'''}$ convergente vers un nombre μ; $0 \leqq \mu \leqq M$ alors que les suites $x'_{i'''}$ et $P(x'_{i'''})$ tendent faiblement vers L_1 et L_2.

On peut par ailleurs remarquer que l'on a:

$$\omega\|\lambda P(x') + Kx'\|_B \leqq \beta[MQ(M) + |K|M].$$

Soit donc L le vecteur de (B) $L = (\mu L_2 + KL_1)\omega$ vers lequel converge faiblement $\lambda\omega P(x'_{i'''}) + K\omega x'_{i'''}$.

Montrons que la transformation $I \times T$ transforme la convergence faible vers (μ, L) en convergence forte vers $(\mu, T \cdot L)$; L à pour composante $L'_{-p}, \cdots, L'_1, \cdots, L'_n$ avec $|p| = 1, 2, \cdots$.

La composante L_0 n'intervient pas ici.

Pour $\omega > \alpha$ et n supérieur à un nombre N_1 constant on a $|t_n| \leqq D/n$, D constante positive, par suite la somme: $S_1 = \sum_n |t_n| \, |V_{ni'''} - L'_n|$ pour $|n| > n_1$ avec $n_1 > N_1$, est inférieure à $S_2 = (2/n_1) \times D \times \sum_{n \geqq n_1} (|V_{ni'''}| + |L'_n|)$ inférieure elle-même à $S_3 = (2D/n_1) \times (MQ(M) + |K|M)$ qui pour n_1 suffisamment grand et supérieur à N_1 peut être rendu inférieur à $\varepsilon/2$. n_1 étant fixé, il suffit de choisir les $x_{i'''}$ pour que chacune des différences $|t_n(V_n - L'_n)|$ soit inférieur à $\varepsilon/4n_1$ avec $n < n_1$.

Puisque pour $n < n_1$, $|t_n| < 1/|K|$ il suffit de prendre les $2n_1$ quantités $\psi_{ni'''}$ avec $\psi_{ni'''} = |V_{ni'''} - L'_n| < \varepsilon K/4n_1$ ce qui est possible, les n_1 étant fixés, les V_n tendant vers les L_n.

Par suite, $\|T(L) - T(x'_{i'''})\| < \varepsilon$ ce qui prouve que la transformation est bien compacte.

(b) Montrons l'existence d'un domaine Ω auquel appartiennent toutes les solutions, cela résulte de § 1 :

(i) $\|x'\|_B \geqq K_2$,

(ii) $\|x'\|_B \leqq K_3/\alpha$ puisque $\omega > \alpha$,

(iii) λ est borné inférieurement par un nombre positif λ_m sinon $\omega \to \omega_1$.

(iv) λ est borné supérieurement par λ_M fini sinon ω n'est pas borné inférieurement et ici $\omega > \alpha$.

On peut définir dans $R \times B$ un domaine Ω auquel appartiennent toutes les solutions, la frontière Ω' n'en contenant aucune avec $(\lambda, x') \in \Omega$, Ω défini par :

$$\frac{\lambda_m}{2} + \frac{K_2}{2} \leqq \|\lambda, x'\| \leqq 2\lambda_M + \frac{2K_3}{\alpha}.$$

Conclusion. ω étant le paramètre, si donc nous admettons l'existence d'un nombre impair de *cycles distincts* pour $\omega_1 - \varepsilon < \beta$, ε petit, nous pouvons affirmer que quelque soit ω il existera toujours au moins une solution (λ, x') ceci en vertu du théorème fondamental de Leray–Schauder en supposant de plus H'_2 vérifié pour la biunivocité.

Considérons maintenant ω non plus comme un paramètre mais comme une inconnue, λ étant alors le paramètre.

Considérons alors l'équation différentielle (E_1) initiale avec λ positif et petit ; nous pouvons pour étudier les solutions (C) appliquer la théorie de Haag [4(b)].

Elle permet d'affirmer que le nombre de cycles est celui des racines positives distinctes de l'équation en ρ :

$$\int_0^{2\pi} P(\rho \sin\theta) \sin\theta \, d\theta = 0.$$

Les ω sont tous voisins de ω_1 avec toujours $\omega < \omega_1$.

Pour λ petit, il y a donc $2p + 1$ branches distinctes de $\omega(\lambda)$ avec $\omega(0) = \omega_1$, par suite il existe alors pour une valeur $\omega = \omega_1 - \varepsilon$, $\varepsilon > 0$ petit, $2p + 1$ valeurs de λ et le même nombre de cycles (C).

La condition d'existence du théorème de Leray–Schauder est vérifiée et l'on peut dès lors affirmer l'existence d'au moins un cycle, (x', λ) quelque soit ω, H'_2 étant supposé vérifiée.

Pour la biunivocité, vérifions maintenant la condition H'_2 pour la transformation :

$$Y = X' - T(X, \Lambda),$$

$$\Lambda = \lambda$$

au voisinage d'une solution ω voisin de ω_1, $|\lambda|$ est donc petit et X' voisin d'une solution sinusoïdale.

Il suffit de montrer que si on a simultanément, η étant petit :

$$\|X'_1 - x'\|_B \leqq \eta,$$

$$\|X'_2 - x'\|_B \leqq \eta,$$

d'où $\|X'_1 - X'_B\| \leqq 2\eta$, l'égalité $X'_1 - T(X'_1, \lambda) = X'_2 - T(X'_2, \lambda)$ entraine alors $X'_1 = X'_2$.

En effet, si $X'_1 - X'_2 = u$, u périodique de période naturelle T vérifie l'équation différentielle :

$$u'' + \omega_1^2 u = \lambda(P(X'_1 + u') - P(X'_1)) = \lambda Q(u),$$

équation différentielle non linéaire, où $|\lambda|$ est petit.

Les solutions périodiques comme u de cette équation s'obtiennent donc par la méthode de Haag–Minorsky et l'on a :

$$u = a \cos \omega t + b \sin \omega t + \gamma(\omega, t)$$

avec $\|\gamma(\omega, t)\|_B$ tendant vers zéro avec $|\omega - \omega_1|$, les valeurs de $\rho = \sqrt{a^2 + b^2}$ sont voisines de celles données par les équations de synchronisation :

$$\int_0^{T_1} Q(x', u') \cos \omega_1 t \, dt = \int_0^{T_1} Q(x', u) \sin \omega_1 t \, dt = 0,$$

les valeurs de ρ sont soient nulles et engendrent $u = 0$, sinon elles sont finies et l'on ne peut alors avoir pour aucune d'entre elles pour η petit :

$$\rho_i \sqrt{2} \leqq \|u_i\|_B \leqq 2\eta.$$

On a donc $u = 0$.

1.3. Étude de l'équation (E') : Retour au paramètre λ. Physiquement le vrai paramètre introduit par les Mécaniciens ou les Electroniciens est λ. Il nous fait donc conclure à partir de celui-ci sur l'existence des cycles.

Montrons que $|\lambda|$ ne peut être borné lorsque ω *tend vers zéro.* Nous allons raisonner par l'absurde :

Supposons donc $|\lambda|$ *borné supérieurement par un nombre* $\lambda_1 > 0$.

Nous avons alors à partir de (E'_1) l'inégalité :

$$\omega_1^2 \int_0^{2\pi} |x| \, dt \leqq \omega^2 \int_0^{2\pi} |x''| \, dt + \lambda_1 \omega \int_0^{2\pi} |P(x')| \, dt.$$

Il en résulte, vu (1.5), (1.8), et $|\lambda| < \lambda_1$, que :

$$\alpha = \frac{1}{2\pi} \int_0^{2\pi} |x| \, dt \leqq L\omega, \qquad L \text{ const.}$$

Multiplions maintenant (E'_1) par x' et intégrons entre t_1 et t_2 avec t_1, $t_2 \in [0, 2\pi]$. On a :

$$\omega^2 \int_{t_1}^{t_2} x'x'' \, dt + \frac{\omega_1^2}{2}(x^2(t_2) - x^2(t_1)) = \lambda\omega \int_{t_1}^{t_2} P(x')x' \, dt.$$

Prenons $|x(t_1)| < \alpha$, vu (1.5), (1.8) et $|\lambda| < \lambda_1$; on en déduit t_2 étant arbitraire :

$$|x(t_2)| \leqq A\sqrt{\omega}, \quad A \text{ const.}, \quad \text{soit } \max|x| \leqq A\sqrt{\omega}.$$

Il en résulte que lorsque ω tend vers zéro $x(t)$ tend vers zéro uniformément.

Soit maintenant $x(t)$ et $u(\tau)$ les deux formes des solutions périodiques de (E_1) et (E'_1) où $x(t)$ est a priori un cycle non nul. On a :

$$(E_1) \qquad x'' + \omega_1^2 x = \lambda P(x'),$$

$$(E'_1) \qquad \omega^2 u'' + \omega_1^2 u = \lambda \omega P(u'),$$

avec $\omega x(t) = u(\tau)$, $x'(t) = u'(\tau)$, $\omega t = \tau$.

Si λ demeure borné lorsque ω tend vers zéro, alors il existe une suite (ω_i, λ_i) telle que ω_i tendant vers zéro, λ_i tend vers λ_0 constante positive.

Les solutions périodiques non nulles existent alors pour ω_i et λ_i.

En vertu des résultats de Poincaré les solutions $x(t)$, $x'(t)$ existent alors pour $\lambda = \lambda_0$ ou disparaissent par paire en fusionnant ; mais les solutions $u(\tau)$, $u'(\tau)$ existent pour $\lambda = \lambda_0$, $\omega = 0$, et $u(\tau)$ nulle est alors l'unique cycle de (E'_1) et a une dérivée $u'(\tau)$ nulle.

On a $u'(\tau) = x'(t) = 0$ alors que $\max|x'(t)| \geqq K_2$, vu l'inégalité (1.6) : il en résulte que $|\lambda|$ ne peut être borné supérieurement par un nombre fini λ_1 lorsque ω tend vers zéro.

Cette remarque n'utilisant pas les propriétés des trajectoires planes est valable pour les cycles dans les espaces de dimension supérieure à 2.

Nous en déduisons donc que ω décroissant de ω_1 à 0, λ prend toutes les valeurs possibles de zéro à l'infini ; ceci en vertu de la conséquence de continuité en théorème de Leray–Schauder.

On peut donc affirmer qu'il existe toujours un cycle non nul quelque soit $\lambda > 0$.

1.4. Comportement de la solution (E_1) lorsque le paramètre λ augmente indéfiniment. Dans le cas présent la période n'est pas 2π, mais T période naturelle de (E_1).

Nous faisons les changements de variables traditionnels en posant $x = \lambda z$, et en prenant un nouveau temps $\tau = \lambda t$.

On a alors le système (E''_1) avec x''^2_t de (E_1) égal à y'_t et l'inégalité (1.10) donnant $\|x''^2_t\| \leqq \omega_1 K_1$ pour la période naturelle T, soit :

$$\frac{dy}{d\tau} = P(y) - \omega_1^2 z, \quad y = x'_t,$$

$$\lambda^2 \frac{dz}{d\tau} = y.$$

La nouvelle période est maintenant $T' = \lambda T$ et par suite la norme de y': $\|y'\|_{H\tau}$ est donnée par:

$$\sqrt{\frac{1}{\lambda T}\int_0^{T'} y_\tau'^2\, d\tau} = \sqrt{\frac{1}{\lambda T}\int_0^{\lambda T}\left(\frac{dy}{dt}\right)^2\left(\frac{dt}{d\tau}\right)^2 d\tau}$$

$$= \sqrt{\frac{1}{\lambda T}\int_0^{T}\left(\frac{dy}{dt}\right)^2\left(\frac{dt}{d\tau}\right) dt}$$

$$= \frac{1}{\lambda}\sqrt{\frac{1}{T}\int_0^{T}\left(\frac{dy}{dt}\right)^2 dt}$$

$$< \frac{K_1\omega_1}{\lambda}$$

conformément à notre conclusion de la page 171.

$\|y'_\tau\|_H$ tend vers zéro avec $1/\lambda$ et par suite $\|P(y) - \omega_1^2 z\|_{H\tau}$ aussi ce qui fournit la forme limite du cycle à un ensemble de valeur de τ de mesure nulle près.

On voit de même que $\|dz/d\tau\|_{H\tau} = (1/\lambda^2)\|y\|_{H\tau}$ tend alors également vers zéro car $\|y\|_{H\tau} \leqq K_1$ ce qui régit le mouvement du point sur le cycle.

1.5. Généralisation. On peut donner deux généralisations directes des résultats précédents:

(i) $P(x')$ peut être changé en $P(x', |x'|)$ polynôme en x' et $|x'|$ dont le terme général est de la forme $a|x'^p|x'^{2q+1}$, alors $P(x', |x'|)\, x''$ est une dérivée exacte.

(ii) Dans $P(x')$, a_1 peut être nul, le terme de plus bas degré étant un terme impair $a_{2K+1}x'^{2K+1}$.

(iii) On peut évidemment combiner les deux cas dans $P(x', |x'|)$, le terme de plus bas degré est alors: $a|x'^p|x'^{2m+1}$, celui de plus haut degré: $b|x'^q|x'^{2n+1}$.

Il est clair que toutes les inégalités utilisées sont maintenues, les démonstrations sont valables dans les trois cas (i), (ii), (iii) indiqués ci-dessus.

2.1. Étude des cycles de l'équation différentielle (E_2). Nous prendrons (E_2) sous la forme (E'_2):

$$x'' + \lambda(P(x') + Q(x)) + \omega_1^2 x = 0.$$

Nous faisons les hypothèses suivantes: λ est un paramètre positif ou nul éventuellement, ω_1 est une constante positive, de plus $P(x')$, $Q(x)$ satisfont aux conditions suivantes (H):

(a) $P(x')$ est un polynôme en x' de degré impair à coefficients constants:

$$P(x') = a_{2n+1}x'^{2n+1} + \cdots + a_1 x'$$

avec $a_1 \neq 0$.

(b) $Q(x)$ est une fonction de x seul, définie quelque soit x; $Q(x)$ est toujours dérivable.

(c) $Q'(x)$ est continue quelque soit x, $Q'(x)$ est négatif sauf pour $x = 0$ où $Q'(0) = 0$; il existe un nombre m constant et négatif avec:

$$Q'(x) \leqq mx^2.$$

La condition (c) entraîne: $Q(x)/x \to -\infty$ si $|x| \to +\infty$.

Pour l'étude du cycle éventuel, nous ramènerons la période à 2π, on pose $\omega t = \tau$, τ est la nouvelle variable et l'on prend pour nouvelle inconnue $u(\tau)$; et en revenant aux notations initiales (x', t) nous avons l'équation (E'):

$$\omega^2 x'' + \lambda\omega\left(P(x') + Q\left(\frac{x}{\omega}\right)\right) + \omega_1^2 x = 0. \tag{E'}$$

Nous allons faire des calculs analogues à ceux de la première partie pour pouvoir appliquer le théorème de Leray–Schauder.

2.2. Étude de la solution éventuelle C (obtention des inégalités fondamentales). On suppose l'existence d'un cycle (C) de période 2π pour (E'):

(i) Par multiplication de (E') par x' et intégration entre 0 et 2π, on a:

$$\int_0^{2\pi} P(x')x'\,dt = 0 \tag{2.1}$$

d'où les mêmes résultats que dans § 1:

$$\|x'\|_H \leqq K_1, \quad \max|x'| \geqq K_2. \tag{2.2}$$

Faisons la même remarque que dans § 1, les résultats (2.2) sont valables pour les solutions de (E) de période inconnue T.

(ii) On multiplie cette fois (E') par x'' et on intègre entre 0 et 2π; en se transformant par intégration par partie:

$$\omega^2 \int_0^{2\pi} x''^2\,dt - \lambda \int_0^{2\pi} Q'\left(\frac{x}{\omega}\right)x'^2\,dt - \omega_1^2 \int_0^{2\pi} x'^2\,dt = 0. \tag{2.3}$$

Il en résulte puisque $Q'(x/\omega) \leqq 0$ que pour $\lambda > 0$,

$$\omega^2 \int_0^{2\pi} x''^2\,dt \leqq \omega_1^2 \int_0^{2\pi} x'^2\,dt \tag{2.4}$$

ce qui entraîne $\omega \leqq \omega_1$.

Réciproquement d'ailleurs si $\omega \geqq \omega_1$ l'égalité (2.3) entraîne alors, $Q'(x)$ étant par hypothèse négatif ou nul, que λ doit être négatif ou nul.

Donc:

$$\text{si} \quad \lambda \geqq 0, \quad \omega \leqq \omega_1,$$
$$\text{si} \quad \omega \geqq \omega_1, \quad \lambda \leqq 0.$$

L'égalité (2.3), jointe à $\lambda > 0$, fournit en outre les inégalités:

$$\|x'\|_H \geqq K_4\omega \quad \text{et} \quad \|x''\|_B \leqq \frac{K_3}{\omega}. \tag{2.5}$$

2.3. Liens entre les paramètres ω, λ et la solution x.

(i) Nous admettrons les résultats de Poincaré et Haag: si $\lambda \to 0$, $\omega \to \omega_1$.

(ii) Étude du cas $\omega \to \omega_1$ et du cas $|\lambda| \to \infty$.

Tout d'abord, nous considérons les expressions:

$$I = \int_0^{2\pi} x^2 x'^2 \, dt$$

et nous voulons démontrer que l'ensemble des I est borné inférieurement pour $\omega > \omega_0$, mais x s'écrit $x_r + \alpha$ avec

$$\alpha = \frac{\lambda\omega}{2\pi\omega_1^2} \int_0^{2\pi} \left(P(x') + Q\left(\frac{x}{\omega}\right) \right) dt \quad \text{et} \quad \int_0^{2\pi} x_r \, dt = 0.$$

α fait intervenir λ, et x dont nous voulons étudier le comportement.

Mais en considérant I comme une fonction du deuxième degré en α, α étant une variable réelle on voit aisément que l'on a:

$$\int_0^{2\pi} x^2 x'^2 \, dt \geqq \frac{\int_0^{2\pi} x'^2 \, dt \cdot \int_0^{2\pi} x_r^2 x'^2 \, dt - \left(\int_0^{2\pi} x_r x_r'^2 \, dt\right)^2}{\int_0^{2\pi} x'^2 \, dt} = \frac{A(x, x')}{\int_0^{2\pi} x'^2 \, dt}.$$

Par suite

$$\int_0^{2\pi} x^2 x'^2 \, dt \geqq \frac{A(x, x')}{2\pi K_1^2};$$

soit N la borne inférieure des $A(x, x')/(2\pi K_1^2)$.

Considérons maintenant l'ensemble des x, x_r pour $\omega \geqq \omega_0$; les x' sont comme auparavant bornées dans leur ensemble et également continues, il en est de même des x_r dont les x' sont les dérivées compte tenu de ce que:

$$\int_0^{2\pi} x_r \, dt = 0.$$

Supposons que N soit nul et ne soit pas atteinte, il existe une suite (x_i', x_{ri}) tel que $A(x, x')$ tende vers zéro, on peut dès lors, vu la remarque précédente, construire une suite d'Arzela-Ascoli $(x_i', x_{ri'})$ convergent vers L', L avec $A(L, L') = 0$.

Mais en vertu de l'inégalité de Schwartz, $A(L, L')$ ne peut être nul que si sur le segment $(0, 2\pi)$ L' et LL' sont proportionnels soit L constant ce qui entrainerait L' nul, or L' est la limite au sens de la continuité uniforme des fonctions x_i', avec $\max |x_i'| \geqq K_2$. N ne peut être nul.

Il en résulte de même que si N est atteint; il ne peut être nul.

Examinons maintenant le cas $\omega \to \omega_1$; on peut appliquer le résultat précédent, l'ensemble des I est borné inférieurement et l'on a à partir de (2.4):

$$(\omega_1^2 - \omega^2) \int_0^{2\pi} x'^2 \, dt \geqq \frac{|\lambda| \, |m|}{\omega_1^2} \int_0^{2\pi} x^2 x'^2 \, dt.$$

Il en résulte que $|\lambda|$ tend vers zéro.

Examinons à partir de

$$\omega_1^2 \int_0^{2\pi} x'^2 \, dt \geqq \frac{|\lambda| \, |m|}{\omega_1^2} \int_0^{2\pi} x^2 x'^2 \, dt.$$

Le cas où $|\lambda|$ *tend vers l'infini*; si $\omega \geqq \omega_0$, le second membre augmente indéfiniment.

Il en résulte que si $|\lambda|$ augmente indéfiniment, ω ne peut être borné inférieurement.

(iii) Montrons maintenant que pour tout segment $\omega_1 - \varepsilon$, ω_0, $\|x\|_B$ est bornée supérieurement; il nous suffit pour cela de montrer que:

$$\alpha = \frac{1}{2\pi}\int_0^{2\pi} x\,dt \quad \text{est borné.}$$

Lorsque $\omega \to 0$, $|\lambda|$ est borné inférieurement par $\lambda_1 > 0$, λ_1 constante.

Pour un cycle on tire de (2.3):

$$|\lambda_1|\left|\int_0^{2\pi} Q'\left(\frac{x}{\omega}\right)x'^2\,dt\right| \leqq \omega_1^2\int_0^{2\pi} x'^2\,dt.$$

Il existe t_0 vu l'hypothèse c sur le signe de $Q'(u)$ tel que:

$$|\lambda_1|\left|Q'\left(\frac{x(t_0)}{\omega}\right)\right|\left|\int_0^{2\pi} x'^2\,dt \leqq \omega_1^2\int_0^{2\pi} x'^2\,dt\right.$$

d'où $|Q'(x(t_0)/\omega)|$ est borné par ω_1^2/λ_1, d'où par suite $|x(t_0)| \leqq F\omega_1$, F constante.

On a:

$$x(t) = x(t_0) + \int_{t_0}^{t} x'\,dt$$

d'où:

$$|x(t)| \leqq F_1\omega_1 + 2\pi K_1$$

d'où $x(t)$ est borné quelque soit $\omega < \omega_1 - \eta$, $0 < \eta < \omega_1$ d'où $\alpha = (1/2\pi)\int_0^{2\pi} x\,dt$ l'est également.

Par suite pour tout intervalle $\omega_1 - \eta$, ω_0, $\|x\|_B \leqq \|x'\|_B + |\alpha|$ est borné supérieurement.

2.4. Existence des solutions. On écrit l'équation sous la forme:

$$\omega^2 X'' + K\omega X' + \omega_1^2 X = K\omega x' - \lambda\omega\left(P(x') + Q\left(\frac{x}{\omega}\right)\right),$$

où K est une constante réelle non nulle fixée une fois pour toute, on prend pour paramètre de la méthode de Leray–Schauder: ω et l'on choisit $\omega_1 - \eta > \omega > \omega_0$ *fixe*. η est positif arbitrairement petit, ω_0 est arbitraire.

Les inconnues sont:

$$\left(X', \alpha = \frac{1}{2\pi}\int_0^{2\pi} X(t)\,dt, \lambda\right).$$

Nous allons procéder de la manière suivante:

(i) Si B est l'espace de Banach des fonctions:

$$x(t) = \sum_{n=-\infty}^{n=+\infty} \alpha_n e^{nit} \quad \text{avec} \quad \|x(t)\|_B = \sum_{n=-\infty}^{n=+\infty} |\alpha_n|,$$

$B' \subset B$ est le sous-espace auquel appartient les $x(t)$ lorsque $\alpha_0 = 0$; c'est également le sous-espace des dérivées:

$$x'(t) = \sum_{n=-\infty}^{n=+\infty} n\alpha_n i\, e^{nit} \quad \text{lorsque} \quad \sum_{n=-\infty}^{n=+\infty} |n\alpha_n| < \infty.$$

Considérons l'espace $B_1 = B' \times R_1$, où R_1 est la droite réelle, $x' = B'$, $\alpha \in R_1$. Nous prendrons dans (B_1) la norme de $(x', \alpha) = \|x'\|_B + |\alpha|$.

x est complètement défini par (x', α).

Nous savons que α est borné par une constante $C > 0$ pour $\omega_0 < \omega < \omega_1 - \eta$.

α appartiendra à la droite R_1, λ est borné pour $\omega_0 < \omega < \omega_1 - \eta$, inférieurement puisque $\omega < \omega_1 - \eta$, supérieurement car $\omega > \omega_0$. λ est alors positif puisque $\omega < \omega_1$.

L'espace $\mathscr{E}$ sera alors $B_1 \times R_1$ avec la norme:

$$\|x'\|_B + |\alpha| + |\lambda|.$$

L'équation s'écrira dans $B_1 \times R_1$ avec $x' = X'$; $\Lambda = \lambda$, $(X', \alpha, \lambda) = T(x', \alpha, \lambda)$ définie par les trois égalités:

(i) $$X' = S(K\omega x' - \lambda\omega\left(P(x') + Q\left(\frac{x}{\omega}\right)\right), \qquad X' \in B',$$

(ii) $$\alpha = \frac{-\lambda\omega}{\omega_1^2} \int_0^{2\pi} \left[P(x') + Q\left(\frac{x}{\omega}\right)\right] dt, \qquad \alpha \in R_1,$$

(iii) $$\Lambda = \lambda \in R_1.$$

La transformation S est une transformation linéaire portant sur:

$$U = \sum_{n=-\infty}^{n=+\infty} U_n e^{nit} = K\omega x' - \lambda\omega\left(P(x') + Q\left(\frac{x}{\omega}\right)\right), \qquad U \in B,$$

$$SU = \sum_{n=-\infty}^{n=+\infty} s_n U_n e^{nit}, \qquad SU \in B',$$

avec

$$s_n = \frac{ni}{\omega_1^2 - n^2\omega^2 + \lambda K n \omega i}.$$

Les conditions d'application du théorème de Leray–Schauder sont vérifiées.

Par suite on prend pour domaine Ω celui défini par les boules de $B_1 \times R_1 \times R_2$:

$$\sum\left(\frac{2K_1}{\omega_0} + C + \lambda_M\right) - \sum\left(\frac{K_2}{2} + \frac{\lambda_m}{2}\right),$$

où $\sum(R)$ est la boule de centre 0 de $B_1 \times R_1 \times R_2$ avec $\|(X', \alpha, \lambda)\| < R$.

Si donc on a démontré, par la méthode de Haag par exemple, l'existence d'un nombre impair de cycles distincts, les ω étant alors voisins de ω_1, mais bien entendu tous plus petits que lui, donc les λ positifs; alors pour tout ω appartenant à l'intervalle ω_1, $\omega_1 - \eta$, η petit, $\eta > 0$ on peut affirmer l'existence d'un nombre impair de cycles distincts, avec $\lambda_1 \cdots \lambda_{2p+1}$ distincts, positifs.

H_2' étant visiblement vérifiée comme dans § 1.2, nous pouvons affirmer alors que pour tout ω appartenant au segment $\omega_1 - \eta$, ω_0 il existe au moins un cycle de pulsation ω, ω_0 constante positive inférieure à $\omega_1 - \eta$.

Donc un cycle existe pour tout ω, $\omega_1 > \omega > 0$; *comme dans* § 1.3 *on voit de plus* λ *tend vers* $+\infty$ *si* $\omega \to 0$.

Il en résulte que pour tout λ *positif il existe au moins un cycle non nul.*

2.5. Formes limites des cycles $\lambda \to \infty$. Dans § 1 nous avons indiqué un procédé permettant de trouver la forme limite des cycles, mais alors dans le second membre de l'équation, x n'apparaissant que sous la forme $Q(x)$ faisons les changements:

$$\lambda z = x, \quad \lambda t = \tau,$$

d'où les équations:

$$y_\tau' = -[P(y) + Q(\lambda z) + \omega_1^2 z],$$
$$\lambda^2 z_\tau' = y.$$

On doit remarquer que:

$$\|y_\tau'\|_{H_\tau} = \frac{1}{\lambda}\|y_t'\|_H \leqq \frac{\omega_1 K_1}{\lambda}$$

d'où:

$$\|P(y) + Q(\lambda z) + \omega_1^2 z\|_{H_\tau} \to 0 \quad \text{avec} \quad 1/\lambda \quad \text{et} \quad \|z_\tau'\|_{H_\tau} \to 0.$$

2.6. Généralisation. Le théorème d'existence précédent peut être étendu aux équations du type:

$$x'' + \lambda(P(x', |x'|) + Q(x)) + \omega_1^2 x = 0,$$

où $P(x', |x'|)$ est polynôme en x' et $|x'|$ dont le terme général est de la forme $a|x'^p|x'^{2q+1}$ et $Q(x)$ une fonction de x satisfaisant aux conditions (b), (c) de § 2.1.

2.7. Application. Krall [5] a rencontré l'équation suivante:

$$x'' + \lambda(x'|x'| - qx' - p^2x^3) + x = 0.$$

Nous allons voir que quelque soit λ cette solution admet un cycle, avec p et q positifs.

Si λ est petit écrivons les équations:

$$x' = y,$$

$$y' = -x + \lambda[y|y| - qy - p^2x^3].$$

Passons en coordonnées polaires ρ, θ: $\rho' = \lambda[\rho^2 \sin\theta|\sin\theta| - q\rho\sin\theta - p^2\rho^3\cos^3\theta]\sin\theta = \lambda F, \theta' = -1 + \lambda[\sin\theta|\sin\theta| - q\sin\theta - p^2\rho^2\cos^2\theta]\cos\theta = \lambda G$.

On pose: $0 = -(1 + \lambda\varepsilon)t + \varphi$, ε constante corrective de la pulsation.

On a un nouveau système:

$$\rho' = \lambda F_1,$$

$$\varphi' = \lambda G_1.$$

On écrit alors les équations de synchronisation en calculant: $\int_0^T F_1\,dt$, $\int_0^T G_1\,dt$, en faisant $T = 2\pi$ et $\lambda = 0$ à l'intérieur de F_1 et de G_1. La première équation fournit ρ, la deuxième fournit ε.

On a ici: $\rho = 3\pi q/8$.

On se trouve ici dans les conditions d'application du théorème général, on peut affirmer l'existence d'un cycle quelque soit $\lambda > 0$.

3. Systèmes d'équations différentielles du 2ème ordre. Nous allons étudier une généralisation de la première partie.

Soit le système:

$$(E_1) \qquad L_1x_1'' + Mx_2'' + K_1x_1 = \lambda P_1(x_1'),$$

$$(E_2) \qquad Mx_1'' + L_2x_2'' + K_2x_2 = \lambda P_2(x_2'),$$

où λ est un paramètre réel.

On fait les hypothèses suivantes:

(i) L_1, L_2, M, K_1, K_2 sont des constantes positives avec $M^2 - L_1L_2 < 0$ (condition F).

(ii) P_1 et P_2 sont des polynômes à coefficients constants tous deux de degrés impairs:

$$P_1(x_1') = a_{2p+1}x_1'^{2p+1} + \cdots + a_1x_1',$$

$$P_2(x_2') = b_{2q+1}x_2'^{2q+1} + \cdots + b_1x_2'.$$

On suppose que a_{2p+1} et b_{2q+1} sont de même signe et que a_1 et b_1 sont des *constantes non nulles de même signe.*

Le système est linéaire lorsque $\lambda = 0$; il admet deux solutions périodiques distinctes de pulsations respectives ω_1 et ω_2 avec $\omega_2 > \omega_1$, celles-ci sont racines de l'équation bicarrée:

$$(K_1 - L_1\omega^2)(K_2 - L_2\omega^2) - M^2\omega^4 = 0.$$

ω_1^2 et ω_2^2 sont également les valeurs stationnaires du rapport:

$$\frac{K_1u^2 + K_2v^2}{L_1u^2 - 2Muv + L_2v^2},$$

où u et v sont des réels quelconques non nuls simultanément.

Nous ramenons la période du cycle à 2π et si ω est la pulsation naturelle du cycle de (E_1) et (E_2) en procédant au retour aux notations en x_1 et x_2; les équations deviennent:

$$(E'_1) \qquad \omega^2(L_1x''_1 + Mx''_2) + K_1x_1 = \lambda\omega P_1(x'_1),$$

$$(E'_2) \qquad \omega^2(Mx''_1 + L_2x''_2) + K_2x_2 = \lambda\omega P_2(x'_2).$$

3.1. Inégalités concernant les normes et les pulsations. Nous allons dans ce qui suit établir des inégalités concernant les normes et les pulsations.

3.1.1. Inégalités relatives aux normes. Etablissons tout d'abord des inégalités concernant les normes. Pour un cycle donné on déduit en multipliant (E'_1) par x'_1, (E'_2) par x'_2 et en intégrant entre 0 et 2π:

$$\int_0^{2\pi} (x'_1P(x'_1) + x'_2P(x'_2))\,dt = 0 \tag{3.1}$$

ce qui fournit l'inégalité:

$$\begin{aligned} &|a_{2p+1}| \int_0^{2\pi} x_1'^{2p+2}\,dt + |(b_{2q+1})| \int_0^{2\pi} x_2'^{2q+2}\,dt \\ &\leqq \sum_{ij} |a_i| \cdot \int_0^{2\pi} |x_1'^{i+1}|\,dt + |b_j| \int_0^{2\pi} |x_2'^{j+1}|\,dt \end{aligned} \tag{3.2}$$

avec $i \neq 2p + 1, j \neq 2q + 1$.

Si nous posons:

$$\frac{1}{2\pi}\int_0^{2\pi} x_1'^{2p+2}\,dt = l^{2p+2},$$

$$\frac{1}{2\pi}\int_0^{2\pi} x_2'^{2q+2}\,dt = m^{2q+2},$$

on déduit de (3.2) $(i \neq 2p + 1, j \neq 2q + 1)$:

$$\underbrace{|a_{2p+1}|l^{2p+2} + |b_{2q+1}|m^{2q+2}}_{A} \leqq \underbrace{\sum_{ij} |a_i|l^{i+1} + |b_j|m^{j+1}}_{B}. \tag{3.3}$$

Ce qui fournit: $H(l, m) < 0$, avec $H(l, m) = A - B$.

Si on considère la courbe du plan (u, v); $H(u, v) = 0$, celle-ci n'ayant aucun point à l'infini, l et m sont bornés supérieurement et l'on a:

$$l \leqq C_1, \quad m \leqq C_1,$$

d'où

$$\alpha_1 = \|x'_1\|_H \leqq C_1, \quad \alpha_2 = \|x'_2\|_H \leqq C_1, \tag{3.4}$$

où C_1 est une constante positive. Une deuxième inégalité peut être déduite de (3.1).

Posons: $\max|x_1'| + \max|x_2'| = M$, et soit m (constante positive) la plus petite des deux valeurs $|a_1|$ et $|b_1|$; on a alors l'inégalité:

$$\sum_i |a_i| M^{i-1} \int_0^{2\pi} x_1'^2\, dt + \sum_j |b_j| M^{j-1} \int_0^{2\pi} x_2'^2\, dt \geqq m \int_0^{2\pi} (x_1'^2 + x_2'^2)\, dt$$

et finalement a fortiori:

$$m \int_0^{2\pi} (x_1'^2 + x_2'^2)\, dt \leqq \left(\sum_{\substack{i \neq 1 \\ j \neq 1}} |a_i| M^{i-1} + \sum |b_j| M^{j-1} \right) \times I,$$

où $I = \int_0^{2\pi} (x_1'^2 + x_2'^2)\, dt$; il en résulte l'inégalité:

$$m \leqq \sum_{i \geqq 2}^{i \leqq 2p+1} |a_i| M^{i-1} + \sum_{j \geqq 2}^{i \leqq 2q+1} |b_j| M^{j-1}. \tag{3.5}$$

M est donc borné inférieurement par un nombre C_2 fixe et positif.

3.1.2. Inégalités relatives aux pulsations. Reprenons maintenant les équations (E_1') et (E_2'), nous obtenons en multipliant (E_1') par x_1'', (E_2') par x_2'' et en intégrant entre 0 et 2π les deux relations et en ajoutant:

$$\omega^2 \left(L_1 \int_0^{2\pi} x''^2\, dt + 2M \int_0^{2\pi} x_1'' x_2''\, dt + L_2 \int_0^{2\pi} x''^2\, dt \right) - K_1 \int_0^{2\pi} x_1'^2\, dt - K_2 \int_0^{2\pi} x_2'^2\, dt = 0. \tag{3.6}$$

Si $\beta_1 = \|x_1''\|_H$, $\beta_2 = \|x_2''\|_H$ on a:

$$\gamma = \frac{1}{2\pi} \int_0^{2\pi} x_1'' x_2''\, dt, \quad |\gamma| \leqq \beta_1 \beta_2$$

(inégalité de Schwartz) ainsi que $\alpha_1 \leqq \beta_1$, $\alpha_2 \leqq \beta_2$.

Or ω^2 est donné par (3.6) soit:

$$\omega^2 = \frac{K_1 \alpha_1^2 + K_2 \alpha_2^2}{L_1 \beta_1^2 + 2M\gamma + L_2 \beta_2^2},$$

ω^2 est donc certainement inférieure ou égale à l'expression:

$$\frac{K_1 \beta_1^2 + K_2 \beta_2^2}{L_1 \beta_1^2 + L_2 \beta_2^2 - 2M \beta_1 \beta_2}. \tag{3.7}$$

(3.6) et (3.7) sont bien entendu toujours définis en raison de l'inégalité $M^2 - L_1 L_2 < 0$ (condition F).

Le maximum de (3.7) n'est autre que ω_2^2, ω_2 plus grande pulsation de régime linéaire de (E_1), (E_2). Par suite si $\lambda \to 0$ et que $\omega \to \omega_2$, il en résulte que:

$$\omega = \omega_2 - K\lambda^2 + \varepsilon\lambda^2 \quad \text{avec} \quad K \geqq 0.$$

3.1.3. Inégalités liant les normes et les pulsations. Nous remarquons qu'il existe deux constantes r, r_1 et r_2 tel que:

$$0 < r_1^2 \leqq Q = \frac{L_1\beta_1^2 - 2M\beta_1\beta_2 + L_2\beta^2}{\beta_1^2 + \beta_2^2} \leqq r_2^2. \tag{3.8}$$

Il en résulte, si on tient compte que (3.6) fournit vu (3.4) l'inégalité:

$$\omega^2(L_1\beta_1^2 - 2M\beta_1\beta_2 + L_2\beta_2^2) \leqq (K_1 + K_2)C_1^2 \tag{3.9}$$

que par suite: $\omega^2 r_1^2(\beta_1^2 + \beta_2^2) \leqq (K_1 + K_2)C_1^2$ et a fortiori:

$$\beta_1 = \|x_1''\|_H \leqq \frac{C_3}{\omega}, \quad \beta_2 = \|x_2''\|_H \leqq \frac{C_3}{\omega}. \tag{3.10}$$

C_3 est une constante positive.

Il en résulte encore:

$$\alpha_1 = \|x_1'\|_B \leqq \frac{C_4}{\omega}, \quad \alpha_2 = \|x_2'\|_B \leqq \frac{C_4}{\omega} \tag{3.11}$$

ainsi que vu (3.5):

$$\|x_1'\|_B + \|x_2'\|_B \geqq C_5, \tag{3.12}$$

où C_3, C_4, C_5 sont des constantes positives.

3.2. Liens entre λ, ω et le cycle (C). Si λ tend vers zéro, nous admettrons les résultats de Malkin [7], ω tend alors vers ω_1 ou ω_2 pulsations des régimes linéaires $\lambda = 0$, $\omega_1 < \omega_2$.

(a) Si $\omega \to \omega_2$ avec $\omega < \omega_2$, λ tend vers zéro; en effet ω_2^2 est le maximum de:

$$\frac{K_1\beta_1^2 + K_2\beta_2^2}{L_1\beta_1^2 + L_2\beta_2^2 - 2M\beta_1\beta_2} = \Omega^2 \tag{3.13}$$

et l'on a:

$$\frac{K_1\alpha_1^2 + K_2\alpha_2^2}{L_1\beta_1^2 + L_2\beta_2^2 + 2M\gamma} = \omega^2 < \Omega^2 < \omega_2^2.$$

Il en résulte d'après (3.8) les résultats suivants:

$$\begin{gathered} 0 \leqq \beta_1^2 - \alpha_1^2 < \varepsilon(\eta), \quad 0 \leqq \beta_2^2 - \alpha_2^2 \leqq \varepsilon(\eta), \\ |\gamma + \beta_1\beta_2| \leqq \varepsilon(\eta) \\ \text{et si} \quad \delta = \frac{L}{2\pi}\int_0^{2\pi} x_1'x_2'\,dt, \quad |\delta + \alpha_1\alpha_2| \leqq \varepsilon(\eta), \end{gathered} \tag{3.14}$$

ε tendant vers zéro avec $\eta = \omega_2 - \omega$.

On peut dès lors montrer que:

$$|\omega^2(L_1x_1'' + Mx_2'') + K_1x_{1r}|_H \quad \text{et} \quad |\omega^2(Mx_1'' + L_2x_2'') + K_2x_{2r}|_H$$

(où x_{1r} et x_{2r} ont la signification du § 1) tendent vers zéro avec η.

On en déduit alors comme dans § 1.1 que λ tend vers zéro.

3.3. Liens entre λ et ω, $|\lambda| \to +\infty$. On peut procéder comme dans § 1.1 par un raisonnement par l'absurde. Supposons $|\lambda| \to \infty$, ω borné inférieurement par ω_0, $\|x_1''\|_H$, $\|x_2''\|_H$ sont bornés supérieurement, les x_1' et x_2' sont donc des fonctions bornées également continues sur 0, 2π. On a de plus deux inégalités:

$$\omega^2(L_1\|x_1''\|_H + M\|x_2''\|_H) + K_1\|x - \frac{1}{2\pi}\int_0^{2\pi} x\,dt\|_H \geqq |\lambda|\omega_0 \times R_1,$$

$$R_1(\text{resp. } R_2) = \left\| P(x_1') - \frac{1}{2\pi}\int_0^{2\pi} P(x_1')\,dt \right\|_H,$$

$$\omega^2(M\|x_1''\|_H + L_2\|x_2''\|_H) + K_2\left\|x - \frac{1}{2\pi}\int_0^{2\pi} x\,dt\right\|_H \geqq |\lambda|\omega_0 \times R_2.$$

Or, on peut donc d'après le théorème d'Ascoli–Arzela construire une suite λ_i, x_{1i}', x_{2i}', où $|\lambda_i| \to \infty$ tels que x_{1i}', x_{2i}' tendent uniformément vers des fonctions continues $L_1(t)$, $L_2(t)$ qui comme précédemment doivent être nulles, ce qui est contradictoire avec: $\max|x_1'| + \max|x_2'| \geqq C_2 > 0$.

3.4. Existence de la solution. On l'écrit sous la forme:

$$(X_1', X_2', \Lambda) = S(x_1', x_2', \lambda).$$

Les notations sont analogues à celles du paragraphe précédent, B_1 l'espace de Banach des séries $x_1 = \sum_{n=-\infty}^{n=+\infty} \alpha_n e^{nit}$ avec $\sum_n |\alpha_n| < \infty$, B_1' le sous-espace de B_1 avec $\alpha_0 = 0$ de même pour x_2, x_2' et B_2', B_2.

On prend $x_1' \in B_1' \subset B_1$, $x_2' \in B_2' \subset B_2$, $\lambda \in R$ droite réelle. La norme dans $\mathscr{E} = B_1' \times B_2' \times R$ est:

$$\|x_1'\|_{B_1'} + \|x_2'\|_{B_2'} + |\lambda|.$$

Les équations définissant S sont:

$$(S_1) \qquad \Lambda = \lambda$$

et les traductions dans B_1' et B_2' du système linéaire:

$$(S_2) \quad \begin{aligned} \omega^2(L_1X_1'' + MX_2'') + A\omega X_1' + K_1X_1 &= \lambda\omega P_1(x_1') + A\omega x_1', \\ \omega^2(MX_1'' + L_2X_2'') + B\omega X_2' + K_2X_2 &= \lambda\omega P_2(x_2') + B\omega x_2', \end{aligned}$$

où A et B sont des constantes réelles non nulles données une fois pour toute.

On doit préciser qu'en explicitant (S_2) on voit que les opérateurs donnant les coefficients des séries de Fourier des X_1', X_2' en fonction de ceux des seconds membres sont parfaitement définis pour un choix particulier de A et B possible lui aussi parce que $L_1L_2 - M^2 > 0$. On suppose A et B ainsi choisies. Pour l'application (L.S) du théorème de Leray–Schauder; l'ensemble Ω se déduit des inégalités (3.11) et (3.12), et du fait que λ est borné supérieurement et inférieurement quand $\omega_0 < \omega < \omega_1 - \eta$.

On vérifierait comme dans § 1.2, que les conditions d'application (H_1) de Leray–Schauder sont vérifiées, ainsi que (H_2').

Toutefois, la question du signe de λ semble faire apparaître une difficulté, puisque λ peut s'annuler pour $\omega = \omega_1 < \omega_2$.

Supposons donc toujours l'existence d'un nombre impair de cycles pour $\omega_2 > \omega > \omega_2 - \eta$, η petit, les λ étant choisis positifs, H'_2 étant vérifiée, faisons décroître ω.

Si ω franchit ω_1, K valeurs de λ distinctes et positives, correspondant à des cycles $C_1 \cdots C_K$ vont devenir négatives après s'être annulées, mais nous pouvons alors faire correspondre aux mêmes cycles K valeurs de λ positives qui demeurent bornées supérieurement positives tant que $\omega > \omega_0 > 0$. La condition de norme de Leray–Schauder est vérifiée.

Il existe alors toujours au moins un cycle; le théorème de Leray–Schauder permet donc d'affirmer alors qu'à tout λ correspond une solution non nulle. Car on peut démontrer comme dans § 1.3 *que* $\lambda \to \infty$ *si* $\omega \to 0$.

3.5. Limite des cycles lorsque $|\lambda| \to \infty$. On voit comme dans la première partie que pour la période naturelle T, on a:

$$\|x_1''\|_H \leqq \frac{\sqrt{K_1 + K_2 C_1}}{r_1}, \quad \|x_2''\|_H \leqq \frac{\sqrt{K_1 + K_2 C_1}}{r_1}.$$

Le système (E_1, E_2) qui devient après passage dans les plans des phases x_1, y_1, x_2, y_2 avec le temps t:

$$L_1 y_1' + M y_2' = \lambda(P_1(y_1)) - K_1 x_1,$$

$$M y_1' + L_2 y_2' = \lambda(P_2(y_2)) - K_2 x_2,$$

se transforme après changement $x_1 = \lambda z_1$, $x_2 = \lambda z_2$, $t = \tau/\lambda$ en:

$$L_1 y_{1\tau}' + M y_{2\tau}' = P_1(y_1) - K_1 z_1, \quad \lambda^2 \frac{dz_1}{d\tau} = y_1,$$

$$M y_{1\tau}' + L_2 y_{2\tau}' = P_1(y_2) - K_2 z_2, \quad \lambda^2 \frac{dz_2}{d\tau} = y_2.$$

Le même calcul que dans la première partie montre que l'on a lorsque $\lambda \to \infty$:

$$\|P_1(y_1) - K_1 z_1\|_{H\tau} \to 0, \quad \left\|\frac{dz_1}{d\tau}\right\|_{H\tau} \to 0,$$

$$\|P_2(y_2) - K_2 z_2\|_{H\tau} \to 0, \quad \left\|\frac{dz_2}{dz}\right\|_{H\tau} \to 0,$$

ce qui donnent l'allure des limites des cycles.

4. Etude d'une équation différentielle d'ordre 3. Nous allons maintenant examiner une équation du 3ème ordre. Pour permettre le développement de l'étude nous la prendrons de la forme (E):

$$y''' + \omega_1^2 y' = \lambda[P(y'') + b_1 y + Q(y')]$$

satisfaisant à:

(i) λ est un paramètre positif; b_1 une constante.

(ii) $P(y'')$ un polynôme de degré impair à coefficients constants:

$$P(y'') = a_{2p+1}y''^{2p+1} + \cdots + a_1 y'';$$

a_{2p+1} est différent de zéro, il en est de même de a_1 et de b_1 qui sont supposés de signe contraire.

(iii) $Q(y')$ est une fonction continuement dérivable. $Q'(y')$ est toujours positif, sauf pour $y' = 0$ ou $Q'(0) = 0$. Il existe de plus un nombre G positif avec $Q'(y') \geqq Gy'^2$.

Nous allons tout d'abord procéder à une étude des cycles éventuels, on ramènera la période T à 2π, en posant $\omega t = \tau$ et on pose $\omega^2 y = Y$ puis on revient à la notation y.

On a dès lors l'équation pour un cycle de pulsation unité:

$$\text{(E)} \qquad \omega^2 y''' + \omega_1^2 y' = \lambda\omega\left[P(y'') + b_1\frac{y}{\omega^2} + Q\left(\frac{y'}{\omega}\right)\right].$$

Etudions maintenant la solution éventuelle (E).

(i) Multiplions par y'' et intégrons entre 0 et 2π, nous avons:

$$\int_0^{2\pi} P(y'')y''\,dt - \frac{b_1}{\omega^2}\int_0^{2\pi} y'^2\,dt = 0.$$

On en déduit deux inégalités:

$$\text{(4.1)} \qquad \begin{aligned} |a_{2n+1}|\int_0^{2\pi} y''^{2n+2}\,dt \leqq |a_{2n}|\int_0^{2\pi}|y''|^{2n+1}\,dt + \sum|a_i|\int_0^{2\pi}|y''|^{i+1}\,dt \\ + \left(|a_1| + \left|\frac{b_1}{\omega^2}\right|\right)\int_0^{2\pi} y''^2\,dt, \qquad i \neq 1, \end{aligned}$$

ce qui fournit une inégalité:

$$l = \sqrt[(2n+2)]{\frac{1}{2\pi}\int_0^{2\pi} y''^{2n+2}\,dt} \leqq K_1(\omega).$$

De même compte tenu de ce que, a_1 et b_1 étant de signes contraires on a en procédant comme dans la première partie; si $M = \max|y''|$,

$$\text{(4.2)} \qquad |a_{2n+1}|M^{2n-1} + |a_{2n}|M^{2n-2} + \cdots + \geqq |a_1|$$

ce qui donne pour M une borne inférieure $K_2 > 0$ indépendante de ω.

(ii) Multiplions (E) par y''' et intégrons entre 0 et 2π, on a:

$$\text{(4.3)} \quad \omega^2\int_0^{2\pi} y'''^2\,dt - \omega_1^2\int_0^{2\pi} y''^2\,dt = \lambda\frac{b_1}{\omega}\int_0^{2\pi} yy'''\,dt + \lambda\omega\int_0^{2\pi} Q\left(\frac{y'}{\omega}\right)y'''\,dt.$$

Vu que:

$$[yy'']_0^{2\pi} = \int_0^{2\pi} y'y''\,dt + \int_0^{2\pi} y'y'''\,dt = 0$$

on a: $\int_0^{2\pi} yy'''\,dt = 0$. En intégrant également par partie on voit que:

$$\lambda\omega\int_0^{2\pi} Q\left(\frac{y'}{\omega}\right)y'''\,dt = -\lambda\int_0^{2\pi} Q'\left(\frac{y'}{\omega}\right)y''^2\,dt$$

qui est négatif. L'égalité (4.3) fournit les inégalités pour $\lambda \geqq 0$:

$$\omega \leqq \omega_1, \tag{4.4}$$

$$\omega^2 \int_0^{2\pi} y'''^2\, dt \leqq \omega_1^2 \int_0^{2\pi} y''^2\, dt, \tag{4.5}$$

réciproquement pour $\omega \geqq \omega_1$ on a $\lambda \leqq 0$.

Les conclusions de cette partie sont analogues à celles des parties précédentes soit pour $\lambda \geqq 0$:

$$\begin{gathered} \|y''\|_H \leqq K_1(\omega), \quad \|y''\|_B \geqq K_2, \quad \|y''\|_B \leqq \frac{K_3}{\omega}, \\ \|y'''\|_H \geqq \frac{K_4\sqrt{3}}{\pi}, \quad \|y''\|_H \geqq K_5\omega, \end{gathered} \tag{4.6}$$

K_2, K_3, K_4, K_5 constantes positives.

4.1. Liens entre ω, λ et la solution. Nous verrions comme antérieurement que le problème du cycle: λ fini, ω tendant vers zéro n'a pas de sens.

(i) Nous admettrons que pour λ tendant vers zéro, tous les cycles s'obtiennent par la méthode de Cesari–Haag [2].

(ii) Nous examinons maintenant le cas $\lambda \to +\infty$. Nous procédons toujours par la même méthode, nous supposons ω borné inférieurement par $\omega_0 < \omega_1$.

L'égalité (4.3):

$$\omega^2 \int_0^{2\pi} y'''^2\, dt - \omega_1^2 \int_0^{2\pi} y''^2\, dt = -\lambda \int_0^{2\pi} Q'\left(\frac{y'}{\omega}\right) y''^2\, dt$$

montre que si $\|y''\|_H$ étant déjà borné par $K_1(\omega_0)$, $\|y'''\|_H$ l'est également. Par suite les fonctions y'' et y' sont bornées dans leur ensemble et également continues quelque soit alors $\lambda > 0$.

On peut donc construire une suite λ_n, y_n', y_n'', ω_n avec $\lambda_n \to \infty$ mais où y_n'', y_n' tendent uniformément vers des fonctions $L''(t)$, $L'(t)$ continues.

La suite peut être choisie de telle manière que ω_n bornée supérieurement et inférieurement tende vers une limite $\omega_0' \geqq \omega_0$. On a:

$$\int_0^{2\pi} Q'\left(\frac{L'}{\omega_0'}\right) L''^2\, dt = 0$$

d'où:

$$G \int_0^{2\pi} \frac{L'^2 L''^2}{\omega_0'^2}\, dt = 0$$

d'où $L'' = 0$ ce qui est absurde, vu les inégalités (4.6).

(iii) Cas où $\omega \to \omega_1$, $\lambda > 0$. On verrait comme dans § 1.1 que λ doit tendre vers zéro.

(iv) On verrait également comme auparavant que ω appartenant à tout segment $\omega_1 - \varepsilon$, ω_0.

$\|y''\|_B$, $\|y'\|_B$ et $\|y\|_B$ sont bornés supérieurement, ceci résulte de ce que $K_1(\omega) \leqq K_1(\omega_0)$.

4.2. Existence de la solution pour $\lambda \to 0$.
On peut comme Cesari poser:

$$y' = x_1,$$
$$\omega_1^{-2} y'' + y = x_2,$$

d'où le système:

$$x_1' + \omega_1^2 x = \overbrace{\lambda(P(x_1') + b_1(x_2 - \omega_1^{-2} x_1') + Q(x_1))}^{\varphi} = \lambda\varphi,$$
$$x_2' = \lambda\omega_1^{-2}\varphi.$$

Il suffit alors de rechercher les solutions périodiques de pulsation voisine de ω_1. Ce problème procède de la méthode de Haag. On vérifierait comme dans § 1.2; pour $|\lambda|$ petit les conditions d'existence des solutions et la biumivocité de la correspondance associée à l'équation, c'est-à-dire (H_2').

4.3. Existence de la solution pour ω quelconque, $\omega < \omega_1$. Pour $\lambda > 0$, nous admettrons l'existence d'un nombre impair de cycles distincts pour $\lambda \to 0$, un nombre impair de solutions (y, λ) s'en déduira pour ω voisin de ω_1 et inférieur à ω_1.

On écrira l'équation sous la forme: $(Y'', \alpha, \Lambda) = T(y'', \alpha, \lambda)$ de la même manière que dans § 3.4 en faisant intervenir une constante K $(K \neq 0)$ permettant d'écrire l'équation sous la forme:

$$\omega^2 Y''' + \omega K Y'' + \omega_1^2 Y' = \lambda\omega\left(P(y'') + b_1 \frac{y}{\omega^2} + Q\left(\frac{y'}{\omega}\right)\right) + K\omega y''$$

et

$$\Lambda = \lambda.$$

Pour tout intervalle $\omega_1 - \varepsilon, \omega_0$ la méthode de Leray–Schauder est applicable et donnera alors un nombre impair de solutions et $\lambda \to \infty$ si $\omega \to 0$.

On peut donc conclure:

Si pour λ petit positif il existe un nombre impair de cycles, elle admet quelque soit λ, au moins un cycle; si $\lambda \to \infty$, ω tend vers zéro.

On peut généraliser pour $P(y'')$ comme pour $P(y')$, dans la première étude, en ce qui concerne les $|y''|$, c'est-à-dire remplacer: $P(y'')$ par $P(y'', \|y''\|) = \sum_{p+q \leqq m} a_{pq} |y''^p| y''^{2q+1}$ avec $a_{2m+1} \neq 0$, $a_{2m+1} = a_{0m} \neq 0$, $a_{01} \neq 0$ et $a_{10} = 0$.

BIBLIOGRAPHIE

[1] L. Bers, *Introduction to Topology*, Courant Institute of Mathematical Sciences, New York, 1955.
[2] L. Cesari, *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Springer-Verlag, Berlin–Heidelberg, 1966.
[3a] R. Faure, *Solutions périodiques de certains systèmes d'équations différentielles non linéaires dépendant d'un paramètre*, J. Math. Pures Appl., 40 (1961), pp. 205–232.
[3b] ———, *Etude de certains systèmes d'équations différentielles à coefficients périodiques, solutions périodiques voisines des positions d'équilibre et applications*, Mémoires et Publications de la Société des Sciences, des Arts et des Lettres du Hainaut, 76 (1962), pp. 141–155.

[4a] J. HAAG, *Les mouvements vibratoires*, Presses Universitaires de France, Paris, 1955.
[4b] ———, *Sur certains systèmes d'équations différentiels à solutions périodiques*, Bull. Sci. Math. (2), 70 (1946), pp. 155–172.
[5] G. KRALL, *Dinamica ed aerodinamica dei fili, III. Problemi non lineari delle vibrazioni visibili*, Atti. Accad. Naz. Lincei. Rend. Cl. Sci. Fis. Mat. Nat. (8), 5 (1948), pp. 197–203.
[6] J. LERAY ET J. SCHAUDER, *Topologie et équations fonctionnelles* Ann. Sci. Ecole Norm. Sup., (1934), pp. 45–78.
[7] I. G. MALKIN, *Some problems in the theory of nonlinear oscillations*, Gostekhisdat, Moscow, 1956. (English translation: AEC-tr-3766 (Book 1) Technical Information Service, United States Atomic Energy Commission.)
[8] H. POINCARÉ, *Les méthodes nouvelles de la mécanique céleste*, Dover, New York, 1957.
[9] R. REISSIG, G. SANSONE ET R. CONTI, *Qualitative Theorie nichtlinearer differentialgleichungen*, Cremonese, Rome, 1963.
[10] ———, *Nichtlinearer differentialgleichungen höherer ordnung*, Cremonese, Rome, 1969.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

TOPOLOGICAL GAMES AND THEIR APPLICATIONS TO PURSUIT PROBLEMS. I*

L. A. PETROSYAN†

Abstract. We consider antagonistic pursuit games with prescribed duration T. The pursuit takes place in an abstract topological space X, on which a distance function ρ is defined. If the pursuer P moves along the trajectory $p(t), 0 \leqq t \leqq T$, and the evader E moves along $e(t), 0 \leqq t \leqq T$, then the payoff is equal to $\rho(p(T), e(T))$.

The information is complete. The evader is discriminated against. We give a sufficient condition for the existence of a solution in the sense of sup inf = inf sup.

In some cases, a solution may be found without the discrimination assumption.

Introduction. The models of pursuit games which use the mathematical apparatus of differential equations are inadequate, because of the fundamental difficulty of describing the class of admissible strategies for the players. Such an opinion was expressed in the works of a number of authors (see [1] and [2]), and was expressed to the author by Professor N. N. Vorob'ev. In [3], pursuit games were studied solely on the basis of topological concepts. In this work, we shall not entirely follow the models proposed in [3] since, in our opinion, introducing multivalued strategies does not entirely correspond to the intuitive idea of the players' behavior in a game. We shall take from [3] the idea of spaces of positions for the pursuer and for the evader. The concept of a strategy will differ from that of [3] and, in our model, strategies will always uniquely determine motions.

1. Description of the spaces of positions of the pursuer P and of the evader E. We shall consider two spaces: $\mathscr{P}$ and $\mathscr{E}$. Points $p \in \mathscr{P}$ represent the positions of the pursuer (the first player), and points $e \in \mathscr{E}$ represent the positions of the evader (the second player). The admissible trajectories of the pursuer are determined by a distance function $d(p_1, p_2)$, where $p_1, p_2 \in \mathscr{P}$, which satisfies the following conditions:

(i) $d(p_1, p_2) > 0$ for $p_1 \neq p_2$, and $d(p_1, p_2) = 0$ for $p_1 = p_2$;

(ii) $d(p_1, p_2) \leqq d(p_1, p_3) + d(p_3, p_2)$;

(iii) for any pair $p_1, p_2 \in \mathscr{P}$, there exists at least one point $p_3 \in \mathscr{P}$ such that $d(p_1, p_3) = d(p_3, p_2) = \frac{1}{2} d(p_1, p_2)$,

(iv) from every infinite sequence p_n, one can choose a convergent subsequence p_{n_i}, i.e., there exists a $\bar{p} \in \mathscr{P}$ such that $\lim_{n \to \infty} d(p_{n_i}, \bar{p}) = 0$.

The pair $(\mathscr{P}, d)$ is said to be a compact space with an unsymmetric, convex distance function.

From condition (iii), we obtain that every two points p_1 and p_2 of $\mathscr{P}$ with $d(p_1, p_2) < \infty$ can be joined by a segment, i.e., there exists a function $p(t), 0 \leqq t \leqq d(p_1, p_2)$, such that $d(p_1, p(t)) = t$ and $d(p(t), p_2) = d(p_1, p_2) - t$ whenever $0 \leqq t \leqq d(p_1, p_2)$.

* Originally published in Vestnik Leningrad Univ., 19 (1969), pp. 55–63. Submitted June 4, 1968. This translation into English has been prepared by K. Makowski.

Translated and printed for this Journal under a grant-in-aid from the National Science Foundation.

† Leningrad State University, Leningrad, USSR.

We shall assume that $\mathscr{E}$ is also a compact space with an unsymmetric convex distance function. We shall use the same letter d to denote the distance function in $\mathscr{E}$. Intuitively, the distance $d(p_1, p_2)(d(e_1, e_2))$ represents the shortest time in which the pursuer (evader) can pass from position p_1 to position p_2 (from e_1 to e_2, respectively).

The trajectories of the pursuer and of the evader are all functions $p(t)$ and $e(t)$, respectively, which are defined for $0 \leqq t \leqq T$ and which satisfy the following conditions: Whenever $0 \leqq t_1 \leqq t_2 \leqq T$, $d(p(t_1), p(t_2)) \leqq t_2 - t_1$ and $d(e(t_1), e(t_2)) \leqq t_2 - t_1$.

It is easy to see that the space of all admissible trajectories of the pursuer and the space of all admissible trajectories of the evader are compact metric spaces in the topology of uniform convergence.

2. Strategy and payoff function in a game which discriminates against the player E for a time $\delta > 0$ into the future.

DEFINITION 1. For any t', $0 \leqq t' \leqq T$, let $a(T - t', p(t'))$ be the set of all possible admissible trajectories $p(t)$ which coincide on the interval $[0, t']$. The set $b(T - t', e(t'))$ is similarly defined.

Let us note that the set $a(T - t, p(t))$ $(b(T - t, e(t)))$ monotonically decreases with time, and that $\bigcap a(T - t, p(t))$ $(\bigcap b(T - e(t)))$ consists of the single unique element $p(t)$ (respectively $e(t)$).

Let $p(t) \in a(T - t', p(t'))$ and $e(t) \in b(T - t', e(t'))$, and let us consider the sets $a(T - (t' + \delta), p(t' + \delta))$ and $b(T - (t' + \delta), e(t' + \delta))$. The family of sets $a(T - (t' + \delta), p(t' + \delta))(b(T - (t' + \delta), e(t' + \delta)))$, for all $p(t) \in a(T - t', p(t'))(e(t) \in b(T - t', e(t')))$, forms a partition of the sets $a(T - t', p(t'))$ $(b(T - t', e(t')))$. We denote this partition by $\mathbf{A}(p(t'), T - t')$ $(\mathbf{B}(e(t'), T - t'))$.

Let $\varphi(p(t), e(t), b(e(t), T - t), T - t)$ be a set-theoretic mapping which assigns to every point $(p(t), e(t), T - t)$ and every set $b(e(t), T - t)$ an element of the partition $\mathbf{A}$. Also, let $\psi(p(t), e(t), T - t)$ be a set-theoretic mapping which assigns to every point $(p(t), e(t), T - t)$ an element of the partition $\mathbf{B}$. The mappings φ and ψ will be said to be strategies of P and E, respectively, if, for $t' > t$, the image $\varphi(p(t'), e(t'), b(e(t'), T - t'), T - t')$ is contained in the image $\varphi(p(t), e(t), b(e(t), T - t), T - t)$ and the image $\psi(p(t'), e(t'), T - t')$ is contained in the image $\psi(p(t), e(t), T - t)$.

The classes of all possible strategies of the players P and E will be denoted by $\mathbf{P}$ and $\mathbf{E}$, respectively.

LEMMA 1. *In every situation (φ, ψ) with fixed initial conditions p_0, e_0, the trajectories of the players' motions are defined as follows:*

$$p(t) = \bigcap_{0 \leqq t \leqq T} \varphi(p(t), e(t), \psi(e(t'), T - t'), T - t),$$

$$e(t) = \bigcap_{0 \leqq t \leqq T} \psi(p(t), e(t), T - t)$$

(for the sake of simplicity, the mappings φ and ψ and the images under these mappings are denoted by the same letters).

To simplify our subsequent presentation, we shall assume that the sets $\mathbf{P}$ and $\mathbf{E}$ coincide for both players. Let $\mathbf{P} = \mathbf{E} = X$. For every pair of points $p \in X$, $e \in X$, we shall define a real-valued function $\rho(p, e)$ which satisfies all the axioms of a metric.

Let $p(t)$, $e(t)$ be a pair of trajectories of the players P and E, starting from initial positions p_0, e_0, in a situation φ, ψ. The game is antagonistic, and the payoff for the player E is

$$K(p_0, e_0; \varphi, \psi) = \rho(p(T), e(T)).$$

The pursuit game thus obtained will be denoted by $\Gamma(p, e, T)$.

Let $C_P^T(p_0)$ denote the set of all points in the space $\mathscr{P}$ which the player P can attain at time T, from the initial position p_0, using all possible strategies $\varphi \in \mathbf{P}$. The set $C_E^T(e_0)$ is similarly defined. We assume that the closures of the sets $C_P^T(p_0)$ and $C_E^T(e_0)$ are compact in the metric ρ. Further, we set

$$\hat{\rho}_T(p_0, e_0) = \sup_{\eta \in C_E^T(e_0)} \inf_{\xi \in C_P^T(p_0)} \rho(\xi, \eta).$$

Let $\hat{\rho}_T(p_0, e_0) = \rho(M_1, M)$ (here M_1 and M belong to the closures of the sets $C_P^T(p_0)$ and $C_E^T(e_0)$). Let us fix an $\varepsilon > 0$. Let us choose a point M_1^ε which belongs to some ε-neighborhood of the point M_1 and to the set $C_P^T(p_0)(M_1^\varepsilon \in D_\varepsilon^P(M_1) \cap C_P^T(p_0))$. Any trajectory $p_2^*(t)$ which joins the points p_0 and $M_1^\varepsilon(p(0) = p_0, p(T) = M_1^\varepsilon)$ will be called a conditionally-ε-optimal trajectory of the pursuer P. A point $M^\varepsilon(M^\varepsilon \in D_\varepsilon^E \cap C_E^T(e_0))$ is similarly determined. Any trajectory $e^*(t)$ which joins the points e_0 and $M^\varepsilon(e(0) = e_0, e(T) = M^\varepsilon)$ will be called a conditionally-ε-optimal trajectory for the player E.

A set of points (p, e) will be said to be a singular surface of the first kind if, for each (p_0, e_0) in the set, there exist two distinct points M' and M'' such that

$$\hat{\rho}_T(p_0, e_0) = \rho(M_1, M') = \rho(M_2, M'').$$

A singular surface will be said to be a dispersal surface (DS) if, for any pair of conditionally-ε-optimal trajectories, the following holds: At some time t, $0 \leqq t \leqq T$, $p^*(t)$, $e^*(t) \in$ (DS) implies that $t = 0$ (i.e., conditionally-ε-optimal trajectories do not intersect a dispersal surface).

A point M will be called a center of pursuit in the game $\Gamma(p_0, e_0, T)$, if

$$\hat{\rho}_T(p_6, e_0) = \rho(M_1, M).$$

A singular surface will be said to be strongly dispersive if no two conditionally-ε-optimal trajectories $e_1^*(t)$, $e_2^*(t)$, which correspond to different centers of pursuit M and M', belong to one element of the partition $\mathbf{B}$.

3. Basic theorems for games with δ-discrimination. A center of pursuit will be said to be ε-invariant if (i) it belongs, in all games of pursuit $\Gamma(p^*(t), e^*(t), T - t)$ (where $p^*(t)$, $e^*(t)$ is any pair of ε-optimal trajectories for the players P and E), to $D_\varepsilon^E(M)$; (ii) there exists a δ, with $0 < \delta < T$, such that, in the games $\Gamma'(p^*(t), e_0, T - t)$ which are obtained from Γ under the additional condition that E rests at e_0 until the time δ, this center belongs to $D_\varepsilon^E(M)$ whenever $0 \leqq t \leqq \delta$. We note that the player P has more information in the game (he knows the choice of the player E at every step), and, thus, the player E turns out to be discriminated against. The following theorem holds.

THEOREM 1. *Let there exist, in a game $\Gamma(p, e, T)$, a unique ε-invariant center of pursuit M, and let $\hat{\rho}_T(p, e) > 0$. Then there exists an ε-saddle point in pure strategies,*

conditionally-ε-optimal trajectories are ε-optimal, and the value of the game is $\hat{\rho}_T(p, e)$.

Proof. From the uniqueness of the point M, it follows that there exists an $\varepsilon > 0$ such that, for all points $\eta \in D_{\varepsilon/2}(M) = \{\xi : \rho(\xi, M) < \varepsilon/2\}$,

$$\rho(\eta, C_P^T(p_0)) > \rho(\eta', C_P^T(p_0)),$$

where η' is an arbitrary point of $\overline{C_E^T(e_0) \setminus D_{\varepsilon/2}(M)}$. Let us now formulate a strategy for the pursuer P which, under the assumptions of the theorem, guarantees a payoff of $\hat{\rho}_T(p_0, e_0) - \varepsilon$. Let E, at every step, choose the set $b(e(t), T - t)$ whose graph has a nonempty intersection with the $\varepsilon/2$-neighborhood $D^{\rho}_{\varepsilon/2}(M)$. Then the pursuer, at every step, chooses the set $a(p(t), T - t) \in \mathbf{A}(p(t), T - t)$, which contains $p_\varepsilon(t)$. Now let E choose, at some time t, the set $b(e(t), T - t) \in \mathbf{B}(e(t), T - t)$ whose graph does not intersect $D^{\rho}_{\varepsilon/2}(M)$, and let M' be a point such that

$$\hat{\rho}_T(p(t), e(t)) = \sup_{\eta \in \bar{b}(e(t), T-t)} \inf_{\xi \in C_P^{T-t}(p(t))} \rho(\xi, \eta) = \rho(M'_1, M')$$

(here, $\bar{b}$ is the graph of the set b). Then the pursuer chooses the set $a(p(t), T - t)$, the closure of whose graph contains the point M'_1.

Let us denote the thus defined strategy of the player P by φ^*_ε. By ψ^*_ε, we shall understand the strategy of the player E such that, at every step, he chooses the set $b(e(t), T - t)$, the closure of whose graph contains the point M.

Let us prove that the situation (φ^*, ψ^*) is an ε-saddle point in the game $\Gamma(p, e, T)$. To do this, it is sufficient to prove that the inequality

$$\varepsilon + K(p_0, e_0; \varphi, \psi^*) \geqq K(p_0, e_0; \varphi^*, \psi^*) \geqq K(p_0, e_0; \varphi^*, \psi) - \varepsilon \tag{1}$$

holds for all $\varphi \in \mathbf{P}$ and all $\psi \in \mathbf{E}$.

Indeed, from the definitions of φ^* and ψ^*, we have

$$\bigcap_{0 \leqq t \leqq T} a(p^*_\varepsilon(t), e^*_\varepsilon(t), T - t, b(e^*_\varepsilon(t), T - t)) = p^*_\varepsilon(t),$$

$$\bigcap_{0 \leqq t \leqq T} b(e^*_\varepsilon(t), p^*_\varepsilon(t), T - t) = e^*_\varepsilon(t)$$

(here, a and b are the behaviors determined by these strategies during the course of the game). Since $p^*_\varepsilon(t)$ and $e^*_\varepsilon(t)$ are conditionally-ε-optimal trajectories,

$$\hat{\rho}_T(p_0, e_0) + \varepsilon \geqq K(p_0, e_0; \varphi^*, \psi^*) \geqq \rho_T(p_0, e_0) - \varepsilon.$$

Now let E employ a strategy $\psi \neq \psi^*$. If ψ coincides with ψ^* along the trajectory $e_\varepsilon(t)$, then the second of inequalities (1) is obvious, since, in this case,

$$K(p_0, e_0; \varphi^*, \psi^*) = K(p_0, e_0; \varphi^*, \psi).$$

Now let t be the first time such that, at the position $e_\varepsilon(t)$, the choice of the player E does not coincide with the choice which is dictated by the strategy ψ^*. Here, two cases are possible: (i) the intersection of $b(e(t), T - t)$ with $D_{\varepsilon/2}(M)$ is nonempty; (ii) this intersection is empty.

In case (i), if $b(e(t), T-t) \cap D_{\varepsilon/2}(M) \neq \varnothing$ for all $t, 0 \leqq t \leqq T$, then the payoff function in the situation (φ^*, ψ) satisfies the inequality

$$\hat{\rho}_T(p_0, e_0) \geqq K(p_0, e_0; \varphi^*, \psi) \geqq \hat{\rho}_T(p_0, e_0) - \varepsilon,$$

and, therefore, the second of inequalities (1) holds.

Now, let t be the time such that $b(e(t), T-t) \cap D_{\varepsilon/2}(M) = \varnothing$ (i.e., case (ii) takes place). Then, since the graph of the set $b(e(t), T-t)$ is contained in the set $C_E^T(e(t), T-t)$, and the center of pursuit is unique,

$$\begin{aligned}\hat{\rho}_T(p_0, e_0) &= \hat{\rho}_{T-t}(p_\varepsilon(t), e_\varepsilon(t)) \\ &= \sup_{\eta \in C_E^T(e_\varepsilon(t),\, T-t)} \inf_{\xi \in C_P^T(p_\varepsilon(t),\, T-t)} \rho(\xi, \eta) \\ &> \sup_{\eta \in b(e(t),\, T-t)} \inf_{\xi \in C_P^T(p_\varepsilon(t),\, T-t)} \rho(\xi, \eta).\end{aligned}$$

Also, the point M' of the set $b(e(t), T-t)$ which is farthest from the set $C_P^T(p_\varepsilon(t), T-t)$ turns out to be strictly closer to the set $C_P^T(p_\varepsilon(t), T-t)$ than the point M. The behavior $a(p(t), T-t)$, which is dictated to the player P by the strategy φ^*, can always guarantee an approach to E to within a distance no greater than

$$\sup_{\eta \in b(e(t),\, T-t)} \inf_{\xi \in C_P^T(p_\varepsilon(t),\, T-t)} \rho(\xi, \eta)$$

(of course, under the assumption that the nonoptimal choice by the player E does not take the game out into the region where the assumptions of the theorem are violated).

The validity of the second of inequalities (1) has been proved.

Now let φ be a strategy for P, and let $p(T) \in C_P^T(p_0)$ be the point at which E arrives in the situation (φ, ψ^*) of the game $\Gamma(p_0, e_0, T)$ at the time T.

Since

$$\hat{\rho}_T(p_0, e_0) = \sup_{\eta \in C_E^T(y)} \inf_{\xi \in C_P^T(x)} \rho(\xi, \eta),$$

then, for any point $\xi \in C_P^T(x)$,

$$\hat{\rho}_T(p_0, e_0) \leqq \rho(\xi, M).$$

In particular,

$$\hat{\rho}_T(p_0, e_0) \leqq \rho(p(T), M),$$

which proves that the first of inequalities (1) holds. The theorem has been proved.

Remark 1. The uniqueness of the center of pursuit implies that $\hat{\rho}_T(p_0, e_0) > 0$ so that Theorem 1 excludes the case of a pointwise capture. Nevertheless, this case is of interest in some applications. In the sequel, we shall return to this case in more detail.

Let us now consider the pursuit game $\Gamma^*(p, e)$ which is dual to the original one (see [7]). The topological structure is the same. The duration is not prescribed in advance. A number $l > 0$ (the capture radius) is given. The aim of player P is to approach E to within a distance which is no greater than l, in minimum time. The player E has the opposite aim. A pursuit problem (a differential-game model) in such a form was formulated by Isaacs [6].

Under certain additional assumptions, one can also successfully apply the method of invariants of pursuit to this problem.

Let $\overline{T}$ be the smallest root of the equation

$$\hat{\rho}_{\overline{T}}(p, e) = l.$$

THEOREM 2. *Let there exist, in a game* $\Gamma(p, e, \overline{T})$ *with duration* $\overline{T}$, *an invariant and unique center of pursuit. Let* $p^*(t), e^*(t)$ *be any pair of optimal trajectories. If* $\rho(p^*(t), e^*(t))$ *is strictly monotonically decreasing with* $t, 0 \leqq t \leqq \overline{T}$, *and if the value of the game* $\Gamma(p^*(t), e^*(t), \overline{T} - t)$ *is monotonically decreasing as a function of time, then the value of the game* $\Gamma^*(p, e)$ *exists and is equal to* $\overline{T}$.

Remark 2. It is easy to show that, instead of monotonicity for $\rho(p^*(t), e^*(t))$, it is sufficient to assume the existence of a conditionally optimal trajectory $e^*(t)$ such that, for all $p^*(t)$ (for all conditionally optimal trajectories of the player P), $\rho(p^*(t), e^*(t)) > l$ for $0 \leqq t \leqq T$.

In case the assumptions of the theorem (the monotonicity) are violated, one can only assert that

$$\overline{T}(p, e) \geqq \operatorname{val} \Gamma^*(p, e).$$

The case where $l = 0$ (the problem was posed in such a way by L. S. Pontryagin in [4]) cannot be immediately obtained from our theorem, since $\hat{\rho}_{\overline{T}}(p, e) = 0$ in this case, and thus there exist infinitely many centers of pursuit (all points of the set $C_E^T(e)$). Nevertheless, a suitable reformulation will lead us to our goal.

We consider the family of games $\Gamma(p, e, T)$ for which $\hat{\rho}_T(p, e) > 0$ and the assumptions of Theorem 2 hold.

Let us denote the center of pursuit in the game $\Gamma(p, e, T)$ by $M(T)$. Let $\overline{T} > 0$ be the first time when $\hat{\rho}_{\overline{T}}(p, e) = 0$ (compare with the first time of absorption in [5]).

DEFINITION 2. If the limit $\lim_{T \to \overline{T}} M(T) = M$ exists, then the point M is said to be a *center of pursuit* in the game

$$\Gamma(p, e, \overline{T}) \quad \text{with} \quad \hat{\rho}_T(p, e) = 0.$$

Making slight changes in the proof of Theorem 2, one can obtain the following theorem.

THEOREM 3. *Let there exist a unique and invariant point* M *in the game* $\Gamma(p, e, \overline{T})$. *Then the value of this game is zero* (*i.e., for any* $\varepsilon > 0$, *the player* P *can assure himself an approach to* E *to within a distance* $\leqq \varepsilon$, *in the time* $\overline{T}$).

COROLLARY. *If* $T \geqq \overline{T}$, *then* P *can passively wait until the time* $T = \overline{T}$, *and then begin his pursuit, assuring thereby a value of the game equal to* $\hat{\rho}_{\overline{T}}(p, e) = 0$ (*of course, if the assumptions of Theorem* 1 *are satisfied in the game from the initial position* $p(T - \overline{T}), e(T - \overline{T})$, *with duration* $\overline{T}$).

Theorems 2 and 3 together yield the following criterion for time-optimal pursuit games (the case of a pointwise capture).

THEOREM 4. *Let* $\overline{T}$ *be the smallest root of the equation* $\hat{\rho}_{\overline{T}}(p, e) = 0$. *Further, suppose that*

$$\lim_{T \to \overline{T}} M(T) = M$$

exists, and assume that, for any pair of conditionally optimal trajectories, the value of the game $\Gamma(p_0, e_0, t)$ *is strictly monotonically decreasing with time. Then the value of the time-optimal pursuit game exists and is equal to* $\overline{T}$. *(The existence of the value of the game is here to be understood in the sense that* sup inf = inf sup.)

4. Global solution for certain games. Let us now consider pursuit games in which all the singular surfaces are strongly dispersive. The theorems which we have presented so far make it possible to construct a solution "locally", i.e., under the additional assumption that the trajectories of the players do not intersect any singular surfaces. To solve a game globally, it is necessary to be able to choose an optimal behavior for the players on the singular surfaces. In the general case, this remains an open problem, but, in the case of a strongly dispersive surface, everything is comparatively simple.

Thus, let p and e be points which belong to a strongly dispersive surface. We shall denote by $\{M\}$ the set of centers of pursuit. Let $M \in \{M\}$ and let $M_1(M)$ be the point in the closure of the set $C_P^T(p)$ which is closest to M. Let $\{M_1\}$ be the set of all such points for different points $M \in \{M\}$. We shall consider conditionally optimal trajectories p_M^* and e_M^* which join the points $p \in M_1(M)$ and $e \in M$. Since the surface is strongly dispersive, the choice made by the player E of an element of $b(e(t), T - t)$ uniquely determines the center of pursuit with respect to the graphs of the sets $b(e(t), T - t)$ and $C_P^T(p)$. Thus, because of the δ-discrimination assumption (P knows the choice made by E), the strategy for P (aimed at the center of pursuit) is uniquely determined once E has made his choice. Making use of Theorems 1–4, one can obtain the following theorem as a corollary.

THEOREM 5. *Suppose that, in a game* $\Gamma(p, e, T)$, *the unique singular surface is strongly dispersive within the entire domain of the game. Then the value of the game is* $\hat{\rho}_T(p, e)$, *an optimal strategy for E consists of aiming at any of the centers of pursuit, and an* ε*-optimal strategy for P consists of aiming at the center of pursuit which is determined by the choice that E makes at each instant of time.*

5. Removing the requirement of discrimination for the time $\delta > 0$. Let us consider the following function. Let $C_P^{T'}(p)$ and $C_E^{T''}(e)$ be the sets of attainability of the players P and E in the times T' and T'', respectively. We shall assume that these sets are closed.

We define

$$\bar{\rho}_{T',T''}(p, e) = \max_{\eta \in C_E^{T''}(e)} \min_{\xi \in C_P^{T'}(p)} \rho(\xi, \eta).$$

Hypothesis. The function $\bar{\rho}_{T'T''}(p, e)$ is continuous in T', T'', p and e.

If this condition is satisfied, then the following theorem holds.

THEOREM 6. *Let there exist, in a game* $\Gamma(p, e, T)$, *a unique and invariant center of pursuit. Then the function* $\hat{\rho}_T(p, e) = \bar{\rho}_{T'T''}(p, e)$ *is the value function of the game without discrimination (i.e., of the game in which the choice made by P is based only on information about the state of the system and the time T).*

Before proving this theorem, let us consider an auxiliary pursuit game $\Gamma'(p, e, T', T)$. The topological structure of the game is the same. The classes of strategies in both games coincide. But the duration of the game is T' for the player

P, and T for the player E, and the payoff function

$$K(p, e; \varphi, \psi) = \rho(p(T'), e(T)).$$

By a simple rephrasing of the proof of Theorem 2 of § 3, we obtain the following result.

THEOREM 7. *Let there exist, in a game* $\Gamma(p, e, T', T)$ *with discrimination* $\delta > 0$, *a unique and invariant center of pursuit. Then the value of the game in pure strategies exists and is equal to* $\bar{\rho}_{T'T}(p, e)$.

We shall now prove the theorem.

Proof. For a given $\varepsilon > 0$, we choose a $\delta > 0$ such that

$$|\rho_T(p_0, e_0) - \bar{\rho}_{T-\delta,T}(p, e)| < \varepsilon \tag{2}$$

for all p and e which belong to $C_P^\delta(p_0)$ and $C_E^\delta(e_0)$.

To prove the theorem, it is sufficient to show that, for any $\varepsilon > 0$, the player P can always guarantee an approach to the player E to within a distance of

$$\rho_{T-\delta,T}(\bar{p}, \bar{e}) = \max_{e \in C_E^\delta(e_0)} \min_{p \in C_P^\delta(p_0)} \bar{\rho}(p, e).$$

But this does indeed take place because, when P arrives from an initial state at a point $p \in C_P^\delta(x_0)$ in the time $\delta > 0$ in an arbitrary way, he already has complete information on the trajectory of E on the interval $(0, \delta)$, since he knows the state $y(\delta)$ (by the hypothesis of the theorem). That is, one can say that the players P and E play the game $\Gamma(p, e, T', T)$ from the states x', y_0, in the times $T' = T - \delta$ and $T = T$, and with discrimination against the player E for a time $\delta > 0$ into the future. Inequality (2) guarantees that an optimal strategy for P in the game $\Gamma(p', e_0, T - \delta, T)$ will assure, for any $\varepsilon > 0$, an approach to within a distance of $\hat{\rho}_T(p'_0, e_0) + \varepsilon$. Obviously, E can also make sure that P will not approach him to within a distance less than $\hat{\rho}_T(p, e)$. This proves the theorem.

The theorem has an obvious application to differential pursuit games. In particular, it allows us to remove the discrimination requirement which was used in an essential way to prove a similar theorem in [7] and [8].

6. Application to games of kind. In [4], L. S. Pontryagin considered, as a basic pursuit problem, the problem of finding the set A of initial positions of the players P and E from which the pursuit can be completed (a pointwise capture can be realized) in the time $T(p_0, e_0)$ under the use of an arbitrary control by the evading player.

It turns out that the preceding results can be applied in order to solve the indicated problem. Namely, the following theorem holds.

THEOREM 8. *For fixed initial conditions* p_0 *and* q_0, *let there exist a nonnegative solution of the equation*

$$\hat{\rho}_T(p_0, q_0) = 0. \tag{3}$$

Also, let there exist, in all games $\Gamma(p_0, q_0, t)$ *with* $0 \leqq t \leqq \overline{T}$ *(here* $\overline{T}$ *is the smallest nonnegative root of* (3)), *a unique and invariant center of pursuit. Then the pair of initial positions* p_0, q_0 *belongs to the set* A *mentioned above.*

The first part of Theorem 8 is a trivial corollary of Theorem 3. The theorem yields an effective algorithm for constructing the set A with the aid of the function $\hat{\rho}_T(p_0, e_0)$.

REFERENCES

[1] L. D. BERKOVITZ, *A variational approach to differential games*, Advances in Game Theory, M. Dresher, L. S. Shapley and A. W. Tucker, eds., Princeton University Press, Princeton, N.J., 1964, pp. 127–174.

[2] W. FLEMING, *The convergence problem for differential games*, Advances in Game Theory, M. Dresher, L. S. Shapley and A. W. Tucker, eds., Princeton University Press, Princeton, N.J., 1964, pp. 195–210.

[3] C. Ryll-Nardzewski, *A theory of pursuit and evasion*, Advances in Game Theory, M. Dresher, L. S. Shapley and A. W. Tucker, eds., Princeton University Press, Princeton, N.J., 1964, pp. 113–126.

[4] L. S. PONTRYAGIN, *On the theory of differential games*, Uspekhi Mat. Nauk, 21 (1966), pp. 219–274.

[5] N. N. KRASOVSKII, *Theory of the Control of Motions*, Nauka, Moscow, 1968.

[6] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[7] L. A. PETROSYAN, *On an invariant in differential pursuit games*, Vestnik Leningrad Univ., 1968, no. 1, pp. 42–51.

[8] ———, *A mapping on a family of differential games of pursuit*, Soviet Math. Dokl., 9 (1968), no. 1, pp. 28–31.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

INPUT-OUTPUT STABILITY OF A BROAD CLASS OF LINEAR TIME-INVARIANT MULTIVARIABLE SYSTEMS*

MATHUKUMALLI VIDYASAGAR†

Abstract. The input-output stability of closed loop control systems, which are not necessarily open loop stable, is considered. The type of stability considered is very broad, and encompasses bounded-input-bounded-output stability. Both continuous-time and discrete-time systems are considered. It is shown that the Desoer–Wu condition [2] is both necessary and sufficient for a large class of closed loop systems to be stable. The criterion is applicable to multivariable systems as well as to single-input-single-output systems.

1. Introduction. One of the best known and most widely used criteria for the stability of closed loop control systems is the Nyquist criterion [1]. The class of systems to which the Nyquist criterion can be applied contains feedback systems with unit feedback gain. The open loop system is assumed to be linear and time invariant, so that the transfer characteristics of the "gain box" are completely specified by a Laplace transform $\hat{g}(s)$. Originally, Nyquist considered scalar gain functions $\hat{g}(s)$ with the property that $\mathscr{L}^{-1}[\hat{g}(s)] = \hat{g}(\cdot) \in L_\infty \cap L_1$, and showed that the closed loop system is stable if and only if $1 + \hat{g}(s)$ had no zeros in the closed half-plane $\operatorname{Re} s \geqq 0$.

It is natural to attempt an extension of Nyquist's results to handle systems which are not open loop stable. Recently, Desoer and Wu [2], [3], and Baker and Vakharia [4] have obtained a sufficient condition for the stability of a large class of such control systems. To describe these results, it is convenient to introduce the sets $\mathscr{A}(\sigma)$ and $\hat{\mathscr{A}}(\sigma)$. The set $\mathscr{A}(\sigma)$ consists of generalized functions of the form

$$g(t) = g_0(t) + \sum_{i=1}^{\infty} g_i \delta(t - t_i), \tag{1}$$

where $g_0(t)$ is a measurable function, δ denotes the unit delta distribution, and it is further true that

$$\int_0^\infty |g_0(t)|\, e^{-\sigma t}\, dt < \infty, \tag{2}$$

$$\sum_{i=1}^{\infty} |g_i|\, e^{-\sigma t_i} < \infty, \tag{3}$$

where σ is a prespecified real number. The set $\mathscr{A}(\sigma)$ becomes a Banach algebra if the norm of an element $g(\cdot) \in \mathscr{A}(\sigma)$ is defined as

$$\|g(\cdot)\| = \int_0^\infty |g_0(t)|\, e^{-\sigma t}\, dt + \sum_{i=1}^{\infty} |g_i|\, e^{-\sigma t_i} \tag{4}$$

and the product of two elements in $\mathscr{A}(\sigma)$ is defined as their convolution. It is easy to verify that $\delta(t)$ is of unit norm and is the unit element for the Banach algebra

* Received by the editors April 19, 1971.

† Department of Electrical Engineering, Sir George Williams University, Montreal, Quebec, Canada. This work was supported by the National Research Council of Canada under Grant A-7790.

$\mathscr{A}(\sigma)$. The set $\hat{\mathscr{A}}(\sigma)$ consists of Laplace transforms of the elements in $\mathscr{A}(\sigma)$. The set $\hat{\mathscr{A}}(\sigma)$ also becomes a Banach algebra if the norm of an element in $\hat{\mathscr{A}}(\sigma)$ is defined as the norm of the corresponding element in $\mathscr{A}(\sigma)$, and the product of two elements in $\hat{\mathscr{A}}(\sigma)$ is defined as their pointwise product. The set $\mathscr{A}^n(\sigma)$ ($\mathscr{A}^{n\times n}(\sigma)$) consists of all vector-valued (matrix-valued) generalized functions, each of whose components belongs to $\mathscr{A}(\sigma)$. The sets $\hat{\mathscr{A}}^n(\sigma)$ and $\hat{\mathscr{A}}^{n\times n}(\sigma)$ are defined analogously.

In [2], Desoer and Wu consider gain functions $\hat{g}(s)$ of the form $\hat{g}(s) = r/(s-\sigma) + \hat{g}_1(s)$, where r is an arbitrary constant and $\hat{g}_1(s) \in \hat{\mathscr{A}}(\sigma)$; in [4], Baker and Vakharia consider gain functions of the form $\hat{g}(s) = \hat{g}_0(s) + \hat{g}_1(s)$, where $\hat{g}_0(s)$ is a rational function and $\hat{g}_1(s) \in \hat{\mathscr{A}}(\sigma)$. In both cases, the authors show that the transfer function of the closed loop system, namely, $\hat{g}(s)/(1+\hat{g}(s))$ is a member of $\hat{\mathscr{A}}(\sigma)$ if the condition

$$\inf_{\operatorname{Re} s \geqq \sigma} |1 + \hat{g}(s)| > 0 \tag{5}$$

holds. In [3], Desoer and Wu consider n-input-n-output multivariable systems where the transfer function matrix $\hat{G}(s)$ is of the form $\hat{G}(s) = R/s + \hat{G}_1(s)$, where R is a type of $n \times n$ matrix and $G_1(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$. They show that the transfer function matrix of the closed loop system is a member of $\hat{\mathscr{A}}^{n\times n}(\sigma)$ if the condition

$$\inf_{\operatorname{Re} s \geqq \sigma} |\det (I + \hat{G}(s))| > 0, \tag{6}$$

holds, where I is the $n \times n$ identity matrix. Finally, Desoer and Vidyasagar [5] show that the conditions (5) and (6) are both necessary and sufficient for closed loop stability. The conditions (5) and (6) are referred to here as the Desoer–Wu conditions.

While the results of [2]–[4] are very useful, it is natural to attempt to expand still further the class of gain functions $\hat{G}(s)$ which are allowed in the forward loop. In this paper, we study the input-output stability of closed loop systems where the forward path transfer function $\hat{G}(s)$ is of the form

$$\hat{G}(s) = \hat{P}(s)\hat{Q}^{-1}(s), \tag{7}$$

where $\hat{P}(s)$ and $\hat{Q}(s)$ belong to $\hat{\mathscr{A}}^{n\times n}(\sigma)$, and the ordered pair $(\det \hat{Q}(s), \det (\hat{P}(s) + \hat{Q}(s)))$ satisfies Condition N, defined below.

Condition N. An ordered pair of scalar-valued functions $(a(s), b(s))$ is said to satisfy Condition N if, whenever $\{s_k\}$ is a sequence in the closed half-plane $\operatorname{Re} s \geqq \sigma$ such that $\lim_{k\to\infty} a(s_k) = 0$, we have $\liminf_{k\to\infty} |b(s_k)| > 0$.

In the case that $a(s)$ and $b(s)$ are meromorphic functions, Condition N amounts to requiring that $a(s)$ and $b(s)$ have no common zeros in the half-plane $\operatorname{Re} s \geqq \sigma$, even at infinity (common poles are allowed, however). This class of gain functions contains the classes considered in [2]–[4] (this is proved later). We show in § 2 that the closed loop transfer function $\hat{H}(s) = \hat{G}(s)(I + \hat{G}(s))^{-1}$ is a member of $\hat{\mathscr{A}}^{n\times n}(\sigma)$ *if and only if* the condition

$$\inf_{\operatorname{Re} s \geqq \sigma} |\det (I + \hat{G}(s))| > 0 \tag{8}$$

is satisfied. This result is extended to discrete-time systems in § 3. Section 4 contains an illustrative example and § 5 comprises the concluding remarks.

2. Continuous-time systems. In this section, the main theorems of the paper are stated and proved.

THEOREM 1. *Suppose a function* $\hat{G}(s)$ *of the form*

$$\hat{G}(s) = \hat{P}(s)\hat{Q}^{-1}(s), \tag{9}$$

where $\hat{P}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$, $\hat{Q}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$, *and the ordered pair* $(\det \hat{Q}(s), \det(\hat{P}(s) + \hat{Q}(s)))$ *satisfies Condition* N, *is given. Then the function* $\hat{H}(s) = \hat{G}(s)(I + \hat{G}(s))^{-1}$ *is a member of* $\hat{\mathscr{A}}^{n\times n}(\sigma)$ *if and only if*

$$\inf_{\operatorname{Re} s \geqq \sigma} |\det(I + \hat{G}(s))| > 0. \tag{10}$$

Proof. Sufficiency of (10). Suppose (10) is satisfied. We wish to show that $\hat{H}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$. We achieve this in two steps: first, we show that (10) implies that

$$\inf_{\operatorname{Re} s \geqq \sigma} |\det(\hat{P}(s) + \hat{Q}(s))| > 0; \tag{11}$$

secondly, we show that (11) implies that $\hat{H}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$. To prove (11), we show that there exists no sequence $\{s_k\}$, with $\operatorname{Re} s_k \geqq \sigma$, such that $|\det(\hat{P}(s_k) + \hat{Q}(s_k))| \to 0$ as $k \to \infty$. Since $\hat{P}(s) + \hat{Q}(s) = (I + \hat{G}(s))\hat{Q}(s)$, it follows that if $\lim_{k\to\infty} |\det(\hat{P}(s_k) + \hat{Q}(s_k))| = 0$, then either $\liminf_{k\to\infty} |\det(I + \hat{G}(s_k))| = 0$ or $\liminf_{k\to\infty} |\det \hat{Q}(s_k)| = 0$. The first possibility is ruled out since (10) is assumed to hold. However, the second possibility can also be ruled out. If $\liminf_{k\to\infty} |\det \hat{Q}(s_k)| = 0$, then there is a subsequence $\{s_{k_i}\}$ of $\{s_k\}$ such that $\lim_{i\to\infty} |\det \hat{Q}(s_{k_i})| = 0$. By Condition N, $\liminf_{i\to\infty} |\det(\hat{P}(s_{k_i}) + \hat{Q}(s_{k_i}))| > 0$, which contradicts the earlier assumption that $|\det(\hat{P}(s_k) + \hat{Q}(s_k))| \to 0$. Hence (11) is proved.

Now we proceed to the second step. It is easy to see that

$$\hat{H}(s) = \hat{P}(s)(\hat{P}(s) + \hat{Q}(s))^{-1}. \tag{12}$$

Let $\hat{D}(s) = \hat{P}(s) + \hat{Q}(s)$; then $D(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ and $\hat{H}(s) = \hat{P}(s)\hat{D}^{-1}(s)$. The matrix $\hat{D}^{-1}(s)$ is given by the cofactor matrix of $\hat{D}(s)$ divided by $\det \hat{D}(s)$. Since sums and products of elements in $\hat{\mathscr{A}}(\sigma)$ once again lie in $\hat{\mathscr{A}}(\sigma)$, the cofactor matrix of $\hat{D}(s)$ is a member of $\hat{\mathscr{A}}^{n\times n}(\sigma)$. Similarly $\det \hat{D}(s) \in \hat{\mathscr{A}}(\sigma)$. So condition (11), together with the results of [6, p. 150], imply that $[1/\det \hat{D}(s)] \in \hat{\mathscr{A}}(\sigma)$, whence $\hat{D}^{-1}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$. This shows that $\hat{H}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$.

Necessity of (10). It is a direct consequence of [5, Theorem 1] that (10) is a necessary condition for $\hat{H}(s)$ to belong to $\hat{\mathscr{A}}^{n\times n}(\sigma)$. Hence the theorem is proved.

Even though Theorem 1 is stated for systems with unity feedback, it can be readily modified to handle systems with any nonsingular constant feedback. Theorem 2 below presents a necessary and sufficient condition for the input-output stability of a class of systems with nonconstant feedback.

THEOREM 2. *Suppose a function* $\hat{G}(s)$ *is of the form*

$$\hat{G}(s) = \hat{P}(s)\hat{Q}^{-1}(s), \tag{13}$$

where $\hat{P}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ *and* $\hat{Q}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$. *Suppose* $\hat{K}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$, *and that the ordered pair* $(\det \hat{Q}(s), \det(\hat{Q}(s) + \hat{K}(s)\hat{P}(s)))$ *satisfies Condition* N. *Then the function*

$\hat{H}(s) = \hat{G}(s)(I + \hat{K}(s)\hat{G}(s))^{-1}$ *is an element of* $\hat{\mathscr{A}}^{n\times n}(\sigma)$ *if and only if*

$$\inf_{\operatorname{Re} s \geqq \sigma} |\det (I + \hat{K}(s)\hat{G}(s))| > 0. \tag{14}$$

Proof. Sufficiency of (14). Suppose (14) holds. Inequality (14) can be equivalently expressed as

$$\inf_{\operatorname{Re} s \geqq \sigma} |\det (I + \hat{K}(s)\hat{P}(s)\hat{Q}^{-1}(s))| > 0. \tag{15}$$

Since $(\det \hat{Q}(s), \det (\hat{Q}(s) + \hat{K}(s)\hat{P}(s)))$ satisfies Condition N, it follows from reasoning analogous to that in the proof of Theorem 1 that (15) implies

$$\inf_{\operatorname{Re} s \geqq \sigma} |\det (\hat{Q}(s) + \hat{K}(s)\hat{P}(s))| > 0. \tag{16}$$

Hence, from [6, p. 150], it follows that $(\hat{Q}(s) + \hat{K}(s)\hat{P}(s))^{-1} \in \hat{\mathscr{A}}^{n\times n}(\sigma)$. Since $\hat{H}(s)$ is also equal to $\hat{P}(s)(\hat{Q}(s) + \hat{K}(s)\hat{P}(s))^{-1}$, we see that $\hat{H}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$. This proves the sufficiency of (14).

Necessity of (14). This part of the proof closely follows that of [5, Theorem 1]. Suppose $\hat{H}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$, i.e., that $\hat{G}(s)(I + \hat{K}(s)\hat{G}(s))^{-1} \in \hat{\mathscr{A}}^{n\times n}(\sigma)$. Since $\hat{K}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$, this implies that $\hat{K}(s)\hat{H}(s) = \hat{K}(s)\hat{G}(s)(I + \hat{K}(s)\hat{G}(s))^{-1} \in \hat{\mathscr{A}}^{n\times n}(\sigma)$. From this it follows that $I - \hat{K}(s)\hat{H}(s) = (I + \hat{K}(s)\hat{G}(s))^{-1} \in \hat{\mathscr{A}}^{n\times n}(\sigma)$, whence $\det (I + \hat{K}(s)\hat{G}(s))^{-1} = 1/\det (I + \hat{K}(s)\hat{G}(s)) \in \hat{\mathscr{A}}(\sigma)$. It can be easily shown that any element in $\hat{\mathscr{A}}(\sigma)$ is bounded over the half-plane $\operatorname{Re} s \geqq \sigma$. Hence the reciprocal of any element in $\hat{\mathscr{A}}(\sigma)$ is bounded away from zero over the half-plane $\operatorname{Re} s \geqq \sigma$, which implies (14). The theorem is proved.

We now proceed to show that the class of systems considered in [2]–[4] is in fact contained in the class of systems covered by Theorems 1 and 2. Consider first of all the scalar case, and suppose $\hat{g}(s)$ is of the form considered by Baker and Vakharia in [4]; namely, let $\hat{g}(s) = \hat{g}_0(s) + \hat{n}(s)/\hat{d}(s)$, where $\hat{g}_0(s) \in \hat{\mathscr{A}}(\sigma)$, and $\hat{n}(s)$ and $\hat{d}(s)$ are polynomials in s with no common factors. To express $\hat{g}(s)$ in the form required by Theorems 1 and 2, let r be the larger of the degrees of $\hat{n}(s)$ and $\hat{d}(s)$, and define $\hat{p}_0(s) = \hat{n}(s)/(s + a)^r$, $\hat{q}_0(s) = \hat{d}(s)/(s + a)^r$, where a is any real number satisfying $a > -\sigma$. It is easy to see that $\hat{p}_0(s) \in \hat{\mathscr{A}}(\sigma)$, $\hat{q}_0(s) \in \hat{\mathscr{A}}(\sigma)$, and that $(\hat{q}_0(s), \hat{p}_0(s))$ satisfies Condition N. Also $\hat{g}(s) = \hat{g}_0(s) + \hat{p}_0(s)/\hat{q}_0(s) = (\hat{g}_0(s)\hat{q}_0(s) + \hat{p}_0(s))/\hat{q}_0(s)$, which is of the form $\hat{p}(s)/\hat{q}(s)$ with $\hat{p}(s) = \hat{g}_0(s)\hat{q}_0(s) + \hat{p}_0(s)$, and $\hat{q}(s) = \hat{q}_0(s)$. It only remains to show that $(\hat{q}(s), \hat{p}(s))$ satisfies Condition N. This is very easy to show, once it is observed that $\hat{g}_0(s)$ is bounded over the half-plane $\operatorname{Re} s \geqq \sigma$.

In [3], Desoer and Wu consider gain functions of the form $\hat{G}(s) = \hat{G}_0(s) + R/(s - \sigma)$, where $\hat{G}_0(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ and R is a nonsingular $n \times n$ matrix. Such a $\hat{G}(s)$ can be expressed as $\hat{P}(s)\hat{Q}^{-1}(s)$, where $\hat{Q}(s) = I \cdot (s - \sigma)/(s + a)$, $\hat{P}(s) = \hat{G}_0(s)Q(s) + R/(s + a)$, and a is any real number satisfying $a > -\sigma$. It is easy to show that the nonsingularity of R implies that $(\det \hat{Q}(s), \det (\hat{P}(s) + \hat{Q}(s)))$ satisfies Condition N. Hence the results contained in Theorems 1 and 2 are actually generalizations of the results in [2]–[4].

In fact, the class of functions studied in [2]–[4] is a proper subset of the class of functions covered by Theorems 1 and 2. For example, consider $\hat{g}(s) = 1/\cosh(s - \sigma)$. Since $\hat{g}(s)$ has an infinite number of poles in the half-plane $\operatorname{Re} s \geqq \sigma$ (namely at

$s = \sigma \pm j(2k + 1)\pi/2$, $k = 0, 1, \cdots, \pm\infty$), it does not fall into the class of functions studied in [2]–[4]. However, if we express $\hat{g}(s)$ as $2e^{-(s-\sigma)}/(1 + e^{-2(s-\sigma)})$, we see that it can be handled by Theorems 1 and 2.

Now we have a remark concerning the application of Theorems 1 and 2 to single-input-single-output systems. If $\hat{g}(s)$, $\hat{p}(s)$ and $\hat{q}(s)$ are all scalar-valued and members of $\hat{\mathscr{A}}(\sigma)$, then the requirement that $(\hat{q}(s), \hat{p}(s) + \hat{q}(s))$ satisfy Condition N can be simplified to: $\hat{p}(s)$ and $\hat{q}(s)$ have no common zeros in the half-plane $\operatorname{Re} s \geqq \sigma$, not even at infinity (note that $\hat{p}(s)$ and $\hat{q}(s)$, being members of $\hat{\mathscr{A}}(\sigma)$, have no singularities in the half-plane $\operatorname{Re} s \geqq \sigma$). Observe, however, that common poles and common zeros in the half-plane $\operatorname{Re} s < \sigma$ are permitted.

3. Discrete-time systems. It is quite clear that analogous versions of Theorems 1 and 2 can be proved for discrete-time systems. In the interest of brevity, the proofs are omitted. Theorem 3 below is a generalization of results due to Desoer and Wu [7], and Desoer and Lam [8].

In what follows, the input and output are both sequences of $n \times 1$ real vectors; l_1 represents the Banach space of absolutely summable sequences, and $l_1^{n\times n}$ is defined in the obvious way; $\mathscr{Z}$ and $\mathscr{Z}^{-1}$ denote z-transformation and inverse z-transformation, respectively; finally "~" denotes z-transformed quantities.

THEOREM 3. *Given a function* $\tilde{G}(z)$ *of the form*

$$\tilde{G}(z) = \tilde{P}(z)\tilde{Q}^{-1}(z), \tag{17}$$

where $\{P_i\} = \mathscr{Z}^{-1}[\tilde{P}(z)]$, $\{Q_i\} = \mathscr{Z}^{-1}[\tilde{Q}(z)]$, $\{P_i\} \in l_1^{n\times n}$, $\{Q_i\} \in l_1^{n\times n}$, *suppose the ordered pair* $(\det \tilde{Q}(z), \det(\tilde{P}(z) + \tilde{Q}(z)))$ *satisfies Condition* N. *Then the inverse z-transform of the function* $\tilde{H}(z) = \tilde{G}(z)(I + \tilde{G}(z))^{-1}$ *is an element of* $l_1^{n\times n}$ *if and only if*

$$\inf_{|z|\geqq 1} |\det(I + \tilde{G}(z))| > 0. \tag{18}$$

THEOREM 4. *Given a function* $\tilde{G}(z)$ *of the form*

$$\tilde{G}(z) = \tilde{P}(z)\tilde{Q}^{-1}(z), \tag{19}$$

where $\{P_i\} \in l_1^{n\times n}$ *and* $\{Q_i\} \in l_1^{n\times n}$, *suppose* $\{K_i\} \in l_1^{n\times n}$ *and that the ordered pair* $(\det \tilde{Q}(z), \det(\tilde{Q}(z) + \tilde{K}(z)\tilde{P}(z)))$ *satisfies Condition* N. *Then the inverse z-transform of the function* $\tilde{H}(z) = \tilde{G}(z)(I + \tilde{K}(z)\tilde{G}(z))^{-1}$ *is an element of* $l_1^{n\times n}$ *if and only if*

$$\inf_{|z|\geqq 1} |\det(I + \tilde{K}(z)\tilde{G}(z))| > 0. \tag{20}$$

4. Example. Consider a feedback system with a forward gain

$$\hat{g}(s) = 1/\cosh s \tag{21}$$

and feedback of constant gain k. Note that $\hat{g}(s)$ is the A-parameter of a uniform LC transmission line, and can also be interpreted as the transfer function of a uniform vibrating string. The gain $\hat{g}(s)$ lies in the class of functions covered by Theorem 1, since $\hat{g}(s) = \hat{p}(s)/\hat{q}(s)$, where

$$\hat{p}(s) = 2e^{-s}, \qquad \hat{q}(s) = 1 + e^{-2s}. \tag{22}$$

Clearly $\hat{p}(s)$ and $\hat{q}(s)$ are members of $\hat{\mathscr{A}}(0)$, and further, whenever $\hat{q}(s) = 0$, we have $p(s) = \pm 2j$, so that Condition N is satisfied. However, $\hat{g}(s)$ has an infinite number

of poles in the half-plane Re $s \geqq 0$, and therefore cannot be handled by any other known methods.

The closed loop system with feedback gain $k \neq 0$ is stable if and only if (cf. Theorem 1).

$$\inf_{\text{Re}\, s \geqq 0} |1 + k\hat{g}(s)| > 0. \tag{23}$$

We have $1 + k\hat{g}(s) = (1 + 2k e^{-s} + e^{-2s})/(1 + e^{-2s})$, so that if $k \neq 0$, then $1 + k\hat{g}(s) = 0$ whenever

$$e^{-s} = -k \pm \sqrt{k^2 - 1}. \tag{24}$$

The quantities $-k \pm \sqrt{k^2 - 1}$ are both real if $|k| \geqq 1$, and are complex if $0 < |k| < 1$. Now, if $0 < |k| < 1$, then both $-k \pm \sqrt{k^2 - 1}$ are complex numbers of magnitude 1, so that all solutions for s of (24) are purely imaginary. In this case (23) does not hold. If $|k| \geqq 1$, then at least one of the quantities $-k \pm \sqrt{k^2 - 1}$ is less than or equal to 1 in magnitude, so that some solutions of (24) satisfy Re $s \geqq 0$. In this case once again (23) does not hold. So the closed loop system is unstable for all nonzero values of k.

5. Conclusions. In this paper, necessary and sufficient conditions have been presented for the input-output stability of a broad class of linear time-invariant systems. An interesting question for future researchers is: what class of Laplace transforms $\hat{g}(s)$ are covered by Theorems 1 and 2? Consider the scalar case, in the interest of simplicity. Then it is clear that any $\hat{g}(s)$ for which Theorem 1 applies necessarily satisfies the following conditions: any singularities of $\hat{g}(s)$ in the half-plane Re $s \geqq \sigma$ must be poles, and these poles must be isolated. However, it is not clear to what extent these conditions are sufficient.

Acknowledgment. Theorem 1 of this paper was inspired by discussions with Professor C. A. Desoer during his visit to Montreal. The author thanks Professor Desoer for his help.

Added in proof. In view of recent results due to Baker and Nasburg [9] and Desoer and Callier [10], Theorem 1 of the present paper can be strengthened to read as follows:

Let G be an $n \times n$ matrix whose elements are Laplace transformable distributions. Then the function $\hat{H}(s) = \hat{G}(s)(I + \hat{G}(s))^{-1}$ is an element of $\hat{\mathscr{A}}^{n \times n}(\sigma)$ if and only if there exist $\hat{P}(s), \hat{Q}(s) \in \hat{\mathscr{A}}^{n \times n}(\sigma)$ such that

(i) $(\det \hat{Q}(s), \det(\hat{P}(s) + \hat{Q}(s)))$ *satisfies Condition* N;

(ii) $\hat{G}(s) = \hat{P}(s)\hat{Q}^{-1}(s)$;

(iii) $\inf_{\text{Re}\, s \geqq \sigma} |\det(I + \hat{G}(s))| > 0$.

REFERENCES

[1] H. Nyquist, *Regeneration theory*, Bell System Tech. J., 2 (1932), pp. 126–147.

[2] C. A. Desoer and M. Y. Wu, *Stability of linear time-invariant systems*, IEEE Trans. Circuit Theory, CT-15 (1968), pp. 245–250.

[3] ———, *Stability of multiple loop feedback linear time-invariant systems*, J. Math. Anal. Appl., 23 (1968), pp. 121–130.

[4] R. A. BAKER AND D. J. VAKHARIA, *Input–output stability of linear time-invariant systems*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 316–319.

[5] C. A. DESOER AND M. VIDYASAGAR, *General necessary conditions for input–output stability*, IEEE Proc., 59 (1971), pp. 1255–1256.

[6] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, American Mathematical Society, Providence, R. I., 1957.

[7] C. A. DESOER AND M. Y. WU, *Input–output properties of discrete systems*, J. Franklin Inst., 290 (1970), pp. 11–24.

[8] C. A. DESOER AND F. L. LAM, *Stability of linear time-invariant discrete systems*, IEEE Proc., 58 (1970), pp. 1841–1843.

[9] R. A. BAKER AND R. E. NASBURG, *Stability of linear time-invariant distributed parameter single loop feedback systems*, Proc. Fifth Asilomar Conference, to appear.

[10] C. A. DESOER AND F. M. CALLIER, *Convolution feedback systems*, Proc. Fifth Asilomar Conference, to appear.

SIAM J. CONTROL
Vol. 10, No. 1, February 1972

NONDISCOUNTED CONTINUOUS TIME MARKOVIAN DECISION PROCESS WITH COUNTABLE STATE SPACE*

PRASADARAO KAKUMANU†

1. Introduction. We are concerned with a continuous time Markov decision process in which the state space $\mathcal{S}$ is countable and the action space $\mathcal{A}$ is finite. The process is observed continuously and found in one of the possible states $i \in \mathcal{S}$; then an action $a \in \mathcal{A}$ is taken; as a result a return $r(i, a)$ is obtained and the process moves to a new state $j \in \mathcal{S}$, which is governed by the transition probability rates $q(j|i, a)$. For each $a \in \mathcal{A}$, let $r(a)$ be the return vector whose ith element is $r(i, a)$, and let $Q(a)$ be the transition rate matrix whose (i, j)th element is $q_{ij}(a) = q(j|i, a), i, j \in \mathcal{S}$.

A deterministic memoryless policy Π is a mapping from $\mathcal{S} \times [0, \infty)$ into $\mathcal{A}$. At any epoch t, if the current state is $x(t) = i$, our action is $a_t = \Pi(i, t)$. We consider only deterministic memoryless policies. In addition, we assume that for every $i \in \mathcal{S}$, $\Pi(i, \cdot)$ is Lebesgue measurable. Such a Lebesgue measurable, memoryless, deterministic policy we call a *Markov policy*. A Markov policy is called stationary if $\Pi(i, t) = \Pi(i)$. That is, a stationary policy Π is a policy which, when the process is in state i, selects action $\Pi(i)$. Let $q_{ij}(t, \Pi) = q(j|i, \Pi(i, t))$ for all $i, j \in \mathcal{S}$ be the transition probability rates from the state i to the state j when policy Π is used. And let $Q(t, \Pi) = \{q_{ij}(t, \Pi); i, j \in \mathcal{S}\}$ be the transition probability rate matrix which we call the *infinitesimal generator* of the Markov decision process which the policy Π is used. If Π is stationary, we write $Q(\Pi)$ instead of $Q(t, \Pi)$. Let $f_{ij}(s, t, \Pi)$, $i, j \in \mathcal{S}$, be the transition probability of going from state i to state j in time $t - s$ when the Markov policy Π is used. For any given Π, let

$$F(s, t, \Pi) = \{f_{ij}(s, t, \Pi); i, j \in \mathcal{S}\} \tag{1.1}$$

be the transition probability matrix corresponding to $Q(t, \Pi)$.

For any two vectors x_1 and x_2, we write $x_1 \geqq x_2$ if the inequality holds for all corresponding coordinates. We call any vector x bounded if $\|x\| = \sup |x_i|$ is bounded. Let e be the infinite column vector with all coordinates unity, and I be the infinite unit matrix.

Let $r(i, t, \Pi) = r(i, \Pi(i, t))$ be the return when the state of the process is $i \in \mathcal{S}$ and the policy Π is used. And let $r(t, \Pi)$ be the return vector. If Π is stationary, we write $r(\Pi)$ instead of $r(t, \Pi)$. For any given Π, the return vector $r(\cdot, \cdot)$, which we assume bounded, is a rate, and the total return over a time interval $[s, t)$ is $\int_s^t r(u, \Pi)\, du$. When policy Π is used the average expected return from the process for each $i \in \mathcal{S}$ is defined to be:

$$\Phi(i, \Pi) = \liminf_{T \to \infty} T^{-1} \int_0^T \sum_j f_{ij}(0, t, \Pi) r(j, t, \Pi)\, dt. \tag{1.2}$$

* Received by the editors May 26, 1970, and in revised form May 27, 1971.

† Department of Industrial Engineering and Operations Research, School of Engineering and Science, New York University, University Heights, Bronx, New York 10453. This work is based on a part of the author's doctoral thesis at Cornell University, Ithaca, New York and was supported by the National Science Foundation under Grant GK 856. Final preparation of this work was done at New York University and was supported by the United States Army Research Office (Durham) in affiliation with the United States Army Munitions Command under Contract DA-31-124-ARO-D-338.

Let $\Phi(\Pi)$ be an infinite column vector with ith element $\Phi(i, \Pi), i \in \mathscr{S}$. For any $\varepsilon > 0$, Π^* will be called *ε-optimal* if for any measurable policy Π, $\Phi(\Pi^*) \geqq \Phi(\Pi) - e\varepsilon$, and will be called *optimal* if it is ε-optimal for every $\varepsilon > 0$, or equivalently, if $\Phi(\Pi^*) \geqq \Phi(\Pi)$.

Howard [3] and Miller [6] have studied the continuous time Markov decision process, when both the state and action spaces are finite. For the expected average return criterion, Derman [2], Ross [7] and Taylor [8] have studied the discrete time Markov decision process when the state space is countable and the action space is finite. In this paper, we investigate the existence of optimal stationary policies when the time parameter is continuous, the state space is countable and the action space is finite.

In § 2, we give sufficient conditions regarding transition probability rates, the Markov process $\{X(t, \Pi), t \geqq 0\}$, and the total expected discount return which imply the existence of an optimal stationary policy for the criterion function (1.1). Under these assumptions in § 3, we show that a bounded set of numbers $\{g, v_j; j \in \mathscr{S}\}$ exists, satisfying the functional equation

$$g = \max_{a \in \mathscr{A}} [r(i, a) + \sum q_{ij}(a) v_j], \qquad i \in \mathscr{S}.$$

Using this functional equation, we shall show the existence of an optimal stationary policy, among the class of Markov policies. Finally, in § 4 under a weaker condition, the existence of ε-optimal stationary policies was shown.

2. Preliminaries and sufficient conditions. In this section, we make certain assumptions regarding the infinitesimal generator, the expected discounted return $\Psi(\Pi, \alpha)$, and the stochastic process $\{X(t, \Pi); t \geqq 0\}$. Using these assumptions in the following sections, we shall show the existence of ε-optimal and optimal stationary policies.

Throughout this paper, we assume that $q_{ij}(t, \Pi)$ satisfies the following assumptions for all $i \in \mathscr{S}$ and $t \geqq 0$.

Assumption 1. $q_{ij}(t, \Pi) \geqq 0$, $i \neq j$, $\sum q_{ij}(t, \Pi) = 0$.

Assumption 2. $|q_{ii}(t, \Pi)| \leqq M$ for some positive number $M < \infty$.

At any epoch t, $\Pi(t)$ specifies some action $a \in \mathscr{A}$. Hence, the above conditions may be stated in terms of $q_{ij}(a)$ for each $a \in \mathscr{A}$ instead of all $t \geqq 0$. Under Assumptions 1 and 2 for a given Π, the author in [4] has shown the existence of a unique stochastic transition probability matrix function $F(s, t, \Pi)$. Further, $F(s, t, \Pi)$ satisfies the Kolmogorov forward differential equations:

$$\frac{\partial}{\partial t} F(s, t, \Pi) = F(s, t, \Pi) Q(t, \Pi) \text{ with } F(s, s, \Pi) = I \tag{2.1}$$

for almost all $t \in [s, \infty)$.

The existence of a measurable Markov process $\{X(t, \Pi); t \geqq 0\}$ corresponding to the stochastic matrix $F(s, t, \Pi)$ was also shown. The Markov process $\{X(t, \Pi); t \geqq 0\}$ is well-behaved and is called the *conservative* Markov process.

We now define the expected discounted criterion function. If the process is initially in state $i \in \mathscr{S}$ and a policy Π is used, the total expected discounted return

to the process with the discount factor $\alpha > 0$ is defined to be:

$$\Psi(i, \Pi, \alpha) = \int_0^\infty e^{-\alpha t} \sum_j f_{ij}(0, t, \Pi) r(j, t, \Pi)\, dt, \qquad i \in \mathscr{S}. \tag{2.2}$$

For this criterion function, the author has shown [5] the existence of an optimal stationary policy, and a procedure to obtain this optimal policy is given. In the sequel, we need the following results, which we give without proof (the reader may refer to Kakumanu [5] for details). If Π is a stationary policy, it was shown that $\Psi(i, \Pi, \alpha)$ is the unique bounded solution to

$$\alpha\Psi(i, \Pi, \alpha) = r(i, \Pi) + \sum q_{ij}(\Pi)\Psi(j, \Pi, \alpha), \qquad i \in \mathscr{S}. \tag{2.3}$$

If Π^* is an optimal stationary policy, then the optimal return $\Psi(i, \Pi^*, \alpha)$, $i \in \mathscr{S}$, is the unique bounded solution of the functional equation:

$$\alpha\Psi(i, \Pi^*, \alpha) = \max_a [r(i, a) + \sum q_{ij}(a)\Psi(j, \Pi^*, \alpha)], \qquad i \in \mathscr{S}. \tag{2.4}$$

Let Π be an α-discounted optimal stationary policy. Define

$$v_{ij}(\Pi, \alpha) = \Psi(i, \Pi, \alpha) - \Psi(j, \Pi, \alpha), \qquad i, j \in \mathscr{S}. \tag{2.5}$$

For the average expected return criterion, when the return rate $r(\cdot, \cdot)$ is bounded and the state space $\mathscr{S}$ is countable, there may not exist an optimal stationary policy. See Derman [2] and Ross [7] for counterexamples. However, it is possible to show the existence of an optimal stationary policy under the following assumption.

Assumption 3. For some sequence $\alpha_m \to 0^+$, and for some state $k \in \mathscr{S}$, there exists a positive constant $M_1 < \infty$ such that $|v_{ik}(\Pi_m, \alpha_m)| \leqq M_1$ for all $m = 1, 2, \cdots$ and $i \in \mathscr{S}$.

Throughout our discussion the Markov decision process starts from the origin. In view of this, we write $F(t, \Pi)$ instead of $F(0, t, \Pi)$, and the (i, j)th element is written as $f_{ij}(t, \Pi)$ instead of $f_{ij}(0, t, \Pi)$ for all $i, j \in \mathscr{S}$. If Π is stationary, the corresponding $F(t, \Pi)$ is a time-homogeneous transition probability matrix function. Since returns are bounded by adding a constant to all the returns, $r(\cdot, \cdot)$ will affect all policies identically for both return functions (1.1) and (2.2); we may without loss of generality assume $r(\cdot, \cdot) \geqq 0$.

Let $T_{ij}(\Pi)$ be the mean first entrance time from state i to j using any stationary policy Π. Let Π_m be the α_m-discounted optimal stationary policy. The set of all stationary policies is a compact set (see Derman [2]). Hence, the limit of any sequence $\{\Pi_m\}$ of stationary policies converges to a stationary policy. The following theorem gives sufficient conditions regarding the Markov process $\{X(t, \Pi); t \geqq 0\}$ which imply Assumption 3.

THEOREM 2.1. *If for some state $k \in \mathscr{S}$ and a sequence $\alpha_m \to 0^+$ there exists a positive constant $M_2 < \infty$ such that $T_{ik}(\Pi_m) \leqq M_2$ for all $i \in \mathscr{S}$ and $m = 1, 2, \cdots$, then the $|v_{ij}(\Pi_m, \alpha_m)|$ are uniformly bounded.*

Proof. Let Π_m be any fixed α_m-discounted optimal policy and suppose the process starts at state $i \in \mathscr{S}$. Let τ be the first passage time from state i to state k.

Using the strong Markov property, we can write the total expected discounted return as

$$\begin{aligned}\Psi(i, \Pi_m, \alpha_m) &= E\int_0^\tau e^{-\alpha_m t} r(\Pi_m)\, dt + E[e^{-\alpha_m \tau}\Psi(k, \Pi_m, \alpha_m)] \\ &\leqq \|r\| E[\tau] + \Psi(k, \Pi_m, \alpha_m) E[e^{-\alpha_m \tau}], \qquad i \in \mathscr{S}.\end{aligned} \tag{2.6}$$

Since $E[\tau] \leqq M_2$ and $E[e^{-\alpha_m \tau}] \leqq 1$, we obtain

$$\Psi(i, \Pi_m, \alpha_m) - \Psi(k, \Pi_m, \alpha_m) \leqq \|r\| M_2, \qquad i \in \mathscr{S}. \tag{2.7}$$

Adding $\Psi(k, \Pi_m, \alpha_m)$ to both sides of (2.6), after simplification we obtain

$$\Psi(k, \Pi_m, \alpha_m) - \Psi(i, \Pi_m, \alpha_m) \leqq (1 - E[e^{-\alpha_m \tau}])\Psi(k, \Pi_m, \alpha_m), \qquad i \in \mathscr{S}.$$

Since $\Psi(k, \Pi_m, \alpha_m) < \alpha_m^{-1}\|r\|$, and $E[e^{-\alpha_m \tau}] \geqq e^{-\alpha_m M}$, we obtain

$$\Psi(k, \Pi_m, \alpha_m) - \Psi(i, \Pi_m, \alpha_m) \leqq (1 - e^{-\alpha_m M_2})\alpha^{-1}\|r\|, \qquad i \in \mathscr{S}. \tag{2.8}$$

Let $M_3 = \|r\| \max [M_1, (1 - e^{-\alpha_m M_2})\alpha^{-1}]$. Then from (2.7) and (2.8) we obtain

$$\|\Psi(k, \Pi_m, \alpha_m) - \Psi(i, \Pi_m, \alpha_m)\| \leqq M_3.$$

From this it is clear that

$$|v_{ij}(\alpha_m)| \leqq 2M_3 < \infty \qquad \text{for all } m = 1, 2, \cdots, \text{ and } i, j \in \mathscr{S}.$$

From this theorem, we can conclude that if the first passage time from any state $i \in \mathscr{S}$ to a particular state $k \in \mathscr{S}$ is bounded for a sequence of Π_m for which $\alpha_m \to 0^+$, then the Assumption 3 will always hold. The following theorem states a condition regarding the transition rates $\{q_{ij}(\Pi)\}$ of the Markov process $\{X(t, \Pi); t \geqq 0\}$ which also implies Assumption 3. In order to prove the theorem, we need the following results which are given in [1].

For any stationary policy, the transition rates $q_{ij}(\Pi)$ are independent of t. When $q_{ij}(\Pi)$ are independent of t, Feller gave an iterative procedure to find the corresponding transition probability matrix $F(t, \Pi)$. The elements $f_{ij}(t, \Pi)$ of the matrix $F(t, \Pi)$ are obtained as a limit of an increasing sequence of functions $f_{ij}^n(t, \Pi)$, which are recursively defined:

$$f_{ij}^0(t, \Pi) = 0,$$

$$f_{ij}^n(t, \Pi) = \delta_{ij} e^{-q_j} + \int_0^t \left[\sum_{l \neq j} f_{il}^{n-1}(t, \Pi) q_{lj}(\Pi)\right] e^{-(t-u)q_j}\, du, \qquad n > 0,$$

where $q_j = -q_{jj}(\Pi)$. Using these equations, we shall prove the following result, which will be used in Theorem 2.3.

LEMMA 2.2. *Let Π be any stationary Markov policy. For any $j \in \mathscr{S}$ if $q_{ij}(\Pi) > 0$ for all $i \in \mathscr{S}, i \neq j$, then $f_{ij}(t, \Pi) > 0$ for all $i \in \mathscr{S}$ and $t > 0$.*

Proof. It was shown by the author in [5] that $f_{ii}(t, \Pi) > 0$ for all $i \in \mathscr{S}$ and $t \geqq 0$ with no positivity restriction on $q_{ij}(\Pi)$. If $i \neq j$, in this case the above recur-

rence relations may be written as:

$$f_{ij}^{n}(t,\Pi) = \int_0^t \left[\sum_{l\neq j} f_{il}^{n-1}(u,\Pi)q_{lj}(\Pi)\right] e^{-(t-u)q_j}\,du$$

$$\geqq \int_0^t f_{ii}^{n-1}(u,\Pi)q_{ij}(\Pi)\,e^{-(t-u)q_j}\,du,$$

because $f_{il}^{n}(u,\Pi) \geqq 0$, $i, l \in \mathcal{S}$, and $q_{lj}(\Pi) \geqq 0$, for $l \neq j$. Hence, if $q_{ij}(\Pi) > 0$, we have $f_{ij}^{n}(t,\Pi) > 0$ for all $t > 0$. Since $f_{ij}(t,\Pi)$ is a limit of the increasing sequence $f_{ij}^{n}(t,\Pi)$, we obtain $f_{ij}(t,\Pi) > 0$.

THEOREM 2.3. *If for some $\beta > 0$ and some state $k \in \mathcal{S}$, $q_{ik}(a) \geqq \beta$ for all $i \in \mathcal{S}$, $i \neq k$, and $a \in \mathcal{A}$, then Assumption 3 holds.*

Proof. Let Π be any stationary policy. Since Π specifies some action $a \in \mathcal{A}$ when the process is in state $i \in \mathcal{S}$, we have by hypothesis $q_{ij}(\Pi) > 0$ for some state $k \in \mathcal{S}$ and for all $i \in \mathcal{S}$. Now from Lemma 2.2 for some $k \in \mathcal{S}$, $f_{ik}(t,\Pi) \geqq \gamma > 0$ for all $t > 0$ and $i \in \mathcal{S}$. It was shown in [1] that the limit of $f_{ij}(t,\Pi)$ exists as $t \to \infty$. For some $k \in \mathcal{S}$, let

$$\beta_{ik}(\Pi) = \lim_{t\to\infty} f_{ij}(t,\Pi), \qquad i \in \mathcal{S}.$$

Hence, for any stationary policy Π, $\beta_{ik}(\Pi) \geqq \gamma$ because $f_{ik}(t,\Pi) \geqq \gamma$ for all $t > 0$. In particular, for any α-discounted optimal stationary policy Π, we have

$$T_{ik}(\Pi) = \frac{1}{\beta_{ik}(\Pi)} \leqq \frac{1}{\gamma} < \infty \qquad \text{for some } k \in \mathcal{S} \text{ and all } i \in \mathcal{S}.$$

The results follow directly from Theorem 2.1.

3. On the existence of stationary optimal policies. In this section, we shall show the existence of optimal stationary policies and also we shall prove they can be obtained from the α-discounted optimal policies. We first show the existence of a bounded set of numbers $\{g, v_j; j \in \mathcal{S}\}$ satisfying the functional equation:

$$g = \max_a \left[r(i,a) + \sum_j q_{ij}(a)v_j\right], \qquad i \in \mathcal{S}. \tag{3.1}$$

Starting from this functional equation, we show the existence of a stationary optimal policy among the class of Markov policies.

THEOREM 3.1. *Under Assumption 3 there exists a bounded set of numbers $\{g, v_j; j \in \mathcal{S}\}$ satisfying*

$$g = \max_a \left[r(i,a) + \sum_j q_{ij}(a)v_j\right], \qquad i \in \mathcal{S}.$$

Proof. Let $\{\alpha_m\}$ be some sequence such that $\alpha_m \to 0^+$ and let Π_m be an α_m-discounted optimal stationary policy. Since $\sum_j q_{ij}(a) = 0$, from (2.4), $\Psi(i,\Pi_m,\alpha_m)$ is the unique bounded solution of

$$\alpha_m\Psi(i,\Pi_m,\alpha_m) = \max_a \left[r(i,a) + \sum_j q_{ij}(a)v_{jk}(\Pi_m,\alpha_m)\right]$$

for some fixed $k \in \mathcal{S}$ and all $m = 1, 2, \cdots$, and $i \in \mathcal{S}$. Since $\mathcal{S}$ is countable, and $|\alpha_m \Psi(i, \Pi_m, \alpha_m)| \leqq \|r\|$, and by Assumption 3, there exists a subsequence $\{\alpha_{m'}\}$ of $\{\alpha_m\}$ tending to 0^+ such that

$$\lim_{m' \to \infty} \alpha_{m'} \Psi(i, \Pi_{m'}, \alpha_{m'}) = g, \qquad i \in \mathcal{S},$$

and

$$\lim_{m' \to \infty} v_{jk}(\Pi_m, \alpha_m) = v_j, \qquad j \in \mathcal{S}.$$

Clearly the numbers $\{g, v_j; j \in \mathcal{S}\}$ are bounded. Taking the limits as $m' \to \infty$ on both sides of (3.1) and since $\mathcal{A}$ is a finite set, we get

$$g = \max_a \left[r(i, a) + \sum_j q_{ij}(a) v_j \right], \qquad i \in \mathcal{S}.$$

LEMMA 3.2. *For any stationary policy* Π, *if there exists a bounded set of numbers* $\{g, v_j; j \in \mathcal{S}\}$, *satisfying*

$$g = r(i, \Pi) + \sum_j q_{ij}(\Pi) v_j, \qquad i \in \mathcal{S}, \tag{3.2}$$

then $g = \Phi(i, \Pi)$, $i \in \mathcal{S}$, *where* Φ *is defined in* (1.1).

Proof. Let $\{f_{li}(t, \Pi); l, i \in \mathcal{S}\}$ be the solution to the forward equations (1.1) when the policy Π is used. Multiplying both sides of (3.2) by $f_{li}(t, \Pi)$ and summing over all $i \in \mathcal{S}$, we get

$$g = \sum_i f_{li}(t, \Pi) r(i, \Pi) + \sum_i f_{li}(t, \Pi) \sum_j q_{ij}(\Pi) v_j, \qquad l \in \mathcal{S}.$$

Since $\sum_i \sum_j |f_{li}(t, \Pi) q_{ij}(\Pi) v_j| < \infty$, the summation signs in the second term of the right-hand side can be interchanged. Using the forward equation (1.1) we obtain

$$g = \sum_j f_{lj}(t, \Pi) r(j, \Pi) + \sum_j \frac{\partial}{\partial t} f_{lj}(t, \Pi) v_j, \qquad l \in \mathcal{S}.$$

Integrating with respect to t from 0 to $T < \infty$, we have

$$\begin{aligned} Tg &= \int_0^T \left[\sum_j f_{lj}(t, \Pi) r(i, \Pi) \right] dt + \int_0^T \left[\sum_j \frac{\partial}{\partial t} f_{lj}(t, \Pi) v_j \right] dt \\ &= \int_0^T \left[\sum_j f_{lj}(t, \Pi) r(i, \Pi) \right] dt + \sum_j f_{lj}(T, \Pi) v_j - v_l, \qquad l \in \mathcal{S}. \end{aligned}$$

Dividing by T and taking the limit as $T \to \infty$, we get

$$g = \lim_{T \to \infty} T^{-1} \int_0^T \sum_i f_{li}(t, \Pi) r(i, \Pi) \, dt = \Phi(l, \Pi), \qquad l \in \mathcal{S}.$$

Thus, g, as defined above, is the expected average return when the policy Π is used. The following theorem establishes the existence of an optimal stationary policy under Assumption 3.

THEOREM 3.3. *If there exists a bounded set of numbers* $\{g, v_j; j \in \mathcal{S}\}$ *satisfying*

$$g = \max_a \left[r(i,a) + \sum_j q_{ij}(a) v_j \right], \qquad i \in \mathcal{S}, \tag{3.1}$$

then there exists a stationary optimal policy Π^* *with* $g = \Phi(i, \Pi^*)$ *for all* $i \in \mathcal{S}$.

Proof. We showed in Theorem 3.1 the existence of a bounded set of numbers $\{g, v_j; j \in \mathcal{S}\}$ satisfying the functional equation (3.1). Let the stationary policy Π^* be defined by $\Pi^*(i) = a_i$ for all $i \in \mathcal{S}$, where $a_i \in \mathcal{A}$ is the decision that maximizes the right-hand side of (3.1). Then

$$g = r(i, \Pi^*) + \sum_j q_{ij}(\Pi^*) v_j, \qquad i \in \mathcal{S}. \tag{3.3}$$

Using Lemma 3.2, we obtain $g = \Phi(i, \Pi^*)$ for all $i \in \mathcal{S}$. Now we shall show that Π^* is an optimal policy over all Markov policies. Let $v(i, t)$ be the expected total return that the system will earn in time t, being started in state $i \in \mathcal{S}$. It can be shown using Theorem 4.2 in [4], that if $v(i, t)$ satisfies the functional equation

$$\frac{dv(i,t)}{dt} = \max [r(i,a) + \sum q_{ij}(a) v(j,t)], \qquad i \in \mathcal{S}, \tag{3.4}$$

for almost all $t \in [0, T]$ and $v(i, T) = 0$, $i \in \mathcal{S}$, then it is the optimal return over the period $[0, T]$ with $v(i, 0) = \max_a r(i, a)$, $i \in \mathcal{S}$. Since $r(i, a)$ and v_i, $i \in \mathcal{S}$ and $a \in \mathcal{A}$, are bounded, we have for some $C < \infty$,

$$v_i - C \leqq v(i,0) \leqq v_i + C, \qquad i \in \mathcal{S}. \tag{3.5}$$

Using (3.4) and (3.5), we shall show for all $i \in \mathcal{S}$ and $\tau \geqq 0$,

$$g\tau + v_i - C \leqq v(i,\tau) \leqq g\tau + v_i + C, \qquad i \in \mathcal{S}. \tag{3.6}$$

Substituting the right-hand inequality in (3.4), we obtain

$$\frac{dv(i,\tau)}{dt} \leqq \max_a \left[r(i,a) + \sum_j q_{ij}(a)(g\tau + v_i + C) \right], \qquad i \in \mathcal{S}.$$

Using (3.1) and $\sum_j q_{ij}(a) = 0$, we obtain

$$\frac{d[v(i,\tau)]}{d\tau} \leqq g.$$

Integrating with respect to τ from 0 to t and using (3.5), we obtain

$$v(i,t) \leqq gt + v_i + C, \qquad i \in \mathcal{S}.$$

Similarly, substituting from the left-hand side of (3.6) in (3.4), we obtain

$$v(i,t) \geqq gt + v_i - C, \qquad i \in \mathcal{S}.$$

The above two relations give (3.6), from which we obtain

$$\lim_{t \to \infty} [v(i,t)/t] = g, \qquad i \in \mathcal{S}.$$

Now let Π be any Markov policy, and $W(i, t)$ be the total expected return from it for any $t > 0$. Since $v(i, t)$ is the optimal return in the period 0 to t, we have

$$\Phi(i, \Pi) = \lim_{t\to\infty} [W(i, t)/t] \leqq \lim_{t\to\infty} [v(i, t)/t] = g, \qquad i \in \mathscr{S}.$$

Hence, Π^* is optimal.

From (3.6), it is clear for some positive $c_1 < \infty$, $v(i, t)$ as a function of time t is bounded by the straight lines $gt + 2c_1$ and $gt - 2c_1$, that is, $|v(i, t) - gt| < 2c_1$, $i \in \mathscr{S}$.

THEOREM 3.4. *If Assumption* 3 *holds, then there exists an optimal stationary policy* Π^* *which is a limit point of* Π_α, *the* α*-discounted optimal policies as* $\alpha \to 0^+$. *Any rule which is a limit point of* $\{\Pi_\alpha\}$ *is optimal.*

Proof. Let $\{\Pi_m\}$ be a sequence of α_m-discounted optimal policies that satisfy Assumption 3. By hypothesis and since the set of stationary policies is compact in the space of all functions from $\mathscr{S}$ to $\mathscr{A}$, there exists a subsequence $\{\Pi_{m'}\}$ of $\{\Pi_m\}$ such that

$$v_{ik}(\alpha_{m'}) \to v_i \quad \text{and} \quad \alpha_{m'}\Psi(i, \Pi_{m'}, \alpha_{m'}) \to g, \qquad i \in \mathscr{S},$$

as $\alpha_{m'} \to 0^+ (m' \to \infty)$ and $\Pi_{m'}$ converges pointwise to some stationary policy Π^*. Since $\alpha_{m'}$ is an $\alpha_{m'}$-discounted optimal policy and $\sum_j q_{ij}(\Pi_{m'}) = 0$ for all $i \in \mathscr{S}$, $\Pi_{m'}$ is chosen such that

$$\alpha_{m'}\Psi(i, \Pi_{m'}, \alpha_{m'}) = r(i, \Pi_{m'}) + \sum_j q_{ij}(\Pi_{m'})v_{jk}(\alpha_{m'}). \tag{3.7}$$

Since $\mathscr{A}$ is finite, $\Pi_{m'} \to \Pi^*$ as $m' \to \infty$ means that for each pair $(i, j) \in \mathscr{S} \times \mathscr{S}$, $q_{ij}(\Pi_{m'}) = q_{ij}(\Pi^*)$ for sufficiently large values of m'. Taking the limits on both sides of (3.7) as $m' \to \infty$ as in Theorem 3.1, we obtain

$$g = r(i, \Pi^*) + \sum_j q_{ij}(\Pi^*)v_j, \qquad i \in \mathscr{S}.$$

From Theorem 3.3 we conclude Π^* is optimal. If $\lim_{m\to\infty} \Pi_m = \Pi$ exists, then as above it can be shown that Π is optimal.

4. ε-optimal policies. Let $\{\alpha_m\}$ be any sequence such that $\alpha_m \to 0^+$ as $m \to \infty$, and let Π_m be an α_m-discounted stationary policy. In this section, we shall show the existence of an ε-optimal stationary policy as a limit of $\{\Pi_m\}$. But the limit of $\{\Pi_m\}$ need not be an ε-optimal policy (see Ross [7] for a counterexample). For a discrete time case, Ross, in [7], showed the existence of ε-optimal stationary policies under certain conditions. In the following theorem, we give a sufficient condition which assures the existence of ε-optimal stationary policies; these conditions are less restrictive than the Assumption 3.

In the following theorem, we shall show for large values of m, Π_m, that the α_m-discounted optimal stationary policy is an ε-optimal policy for the average expected criterion function (1.1). That is, for $\varepsilon > 0$ there exists a positive number $N < \infty$ such that

$$\Phi(i, \Pi_m) \geqq \Phi(i, \Pi) - \varepsilon, \qquad i \in \mathscr{S},$$

for all $m \geqq N$, and for any Markov policy Π.

THEOREM 4.1. *If for some sequence* $\alpha_m \to 0^+$ *there exists* $M_i < \infty$ *for each* $i \in \mathcal{S}$ *such that*

$$\Psi(i, \Pi_m, \alpha_m) - \Psi(j, \Pi_m, \alpha_m) \leqq M_i \tag{4.1}$$

for all $m = 1, 2, \cdots$, *and for all* $i, j \in \mathcal{S}$, *then for large* m, Π_m *is* ε*-optimal.*

Proof. Since Π_m is a stationary policy we have from (2.3):

$$\alpha_m \Psi(j, \Pi_m, \alpha_m) = r(j, \Pi_m) + \sum_k q_{ik}(\Pi_m)\Psi(k, \Pi_m, \alpha_m), \qquad j \in \mathcal{S}.$$

Multiplying by $f_{li}(t, \Pi_m)$ on both sides and summing over all $j \in \mathcal{S}$, we obtain

$$\alpha_m \sum_j f_{lj}(t, \Pi_m)\Psi(j, \Pi_m, \alpha_m) = \sum_j f_{lj}(t, \Pi_m) r(j, \Pi_m) + \sum_j f_{lj}(t, \Pi_m) \sum_k q_{jk}(\Pi_m)\Psi(k, \Pi_m, \alpha_m);$$

using (2.1), we obtain

$$\alpha_m \sum_j f_{lj}(t, \Pi_m)\Psi(j, \Pi_m, \alpha_m) = \sum_j f_{lj}(t, \Pi_m) r(j, \Pi_m) + \sum_k \frac{d}{dt} f_{lk}(t, \Pi)\Psi(k, \Pi_m, \alpha_m), \qquad l \in \mathcal{S}.$$

Using (4.1), we obtain

$$\sum_j f_{lj}(t, \Pi_m) r(j, \Pi_m) \geqq \alpha_m \Psi(i, \Pi_m, \alpha_m) - \alpha_m M_i - \sum_k \frac{d}{dt} f_{lk}(t, \Pi_m)\Psi(k, \Pi_m, \alpha_m).$$

Now integrating over t from 0 to T and dividing by T, then taking limits as $T \to \infty$, we obtain

$$\Phi(l, \Pi_m) \geqq \alpha_m \Psi(i, \Pi_m, \alpha_m) - \alpha_m M_i, \qquad i \in \mathcal{S},$$

for all $i, l \in \mathcal{S}$. Since $\alpha_m \to 0$ as $m \to \infty$, we have

$$\liminf_{m \to \infty} \Phi(l, \Pi_m) \geqq \liminf_{m \to \infty} \alpha_m \Psi(i, \Pi_m, \alpha_m) \tag{4.2}$$

for all $i, l \in \mathcal{S}$. Since the stationary policy Π_m is an α_m-discounted optimal policy, we have for all $i \in \mathcal{S}$ and for any Markov policy Π,

$$\alpha_m \Psi(i, \Pi_m, \alpha_m) \geqq \alpha_m \Psi(i, \Pi, \alpha_m).$$

From this we have

$$\begin{aligned} &\liminf_{m \to \infty} \alpha_m \Psi(i, \Pi_m, \alpha_m) \\ &\quad \geqq \liminf_{m \to \infty} \alpha_m \Psi(i, \Pi, \alpha_m) \\ &\quad \geqq \liminf_{T \to \infty} T^{-1} \int_0^T \left[\sum_j f_{ij}(t, \Pi)\, r(j, \Pi) \right] dt = \Phi(i, \Pi), \qquad i \in \mathcal{S}. \end{aligned}$$

The last inequality follows from Theorem 1 of Widder [8, p. 181]. Hence, we have from (4.2), for any Markov policy Π:

$$\liminf_{m\to\infty} \Phi(l, \Pi_m) \geqq \Phi(i, \Pi), \qquad i, l \in \mathscr{S}.$$

Let

$$g(i) = \sup_{\Pi} \Phi(i, \Pi), \qquad i \in \mathscr{S};$$

then

$$\liminf_{m\to\infty} \Phi(l, \Pi_m) \geqq g(i), \qquad i, l \in \mathscr{S}. \tag{4.3}$$

Since (4.3) is true for all $l \in \mathscr{S}$, putting $i = l$ we obtain

$$\lim_{m\to\infty} \Phi(l, \Pi_m) = g(l), \qquad l \in \mathscr{S}. \tag{4.4}$$

(In the following Lemma we shall show that this limit is uniform in $i \in \mathscr{S}$.) From (4.3) and (4.4), we obtain

$$g(l) \geqq g(i), \qquad i, l \in \mathscr{S}. \tag{4.5}$$

Hence, from (4.4) and (4.5), we obtain

$$\lim_{m\to\infty} \Phi(i, \Pi_m) = g, \qquad i \in \mathscr{S},$$

which shows that for large m, Π_m is ε-optimal.

LEMMA 4.2. *Under the conditions of Theorem* 4.1, $\Phi(i, \Pi_m)$ *converges uniformly in* $i \in \mathscr{S}$ *as* $m \to \infty$.

Proof. Let $i = k$, some fixed state. From Theorem 4.1 we have

$$\liminf_{m\to\infty} \alpha_m \Psi(k, \Pi_m, \alpha_m) \geqq g(k)$$

and

$$\Phi(l, \Pi_m) \geqq \alpha_m \Psi(k, \Pi_m, \alpha_m) - \alpha_m M_k$$

for any state $l \in \mathscr{S}$. For $\varepsilon > 0$, let m_0 be such that $m > m_0$ implies

$$\alpha_m M_k \leqq \varepsilon/2 \quad \text{and} \quad \alpha_m \Psi(k, \Pi_m, \alpha_m) \geqq g(k) - \varepsilon/2.$$

Hence, we have for $m > m_0$,

$$\Phi(l, \Pi_m) \geqq g(k) - \varepsilon, \qquad l \in \mathscr{S}.$$

Using Theorem 4.1 the lemma follows.

5. Acknowledgment. The author acknowledges a great debt of gratitude to his thesis advisor, Professor Howard M. Taylor III, for his guidance and encouragement during the preparation of the thesis on which this paper is based. He is also indebted to the referee for his helpful comments and suggestions.

REFERENCES

[1] K. L. CHUNG, *Markov Chains with Stationary Transition Probabilities*, 2nd ed., Springer-Verlag, Berlin, 1967.

[2] C. DERMAN, *Denumerable state Markovian decision process-average cost criterion*, Ann. Math. Statist., 37 (1966), pp. 1545–1553.

[3] R. A. HOWARD, *Dynamic Programming and Markov Processes*, John Wiley, New York, 1960.

[4] P. KAKUMANU, *Continuous time Markov decision models with applications to optimization problems*, Tech. Rep. 63, Cornell University, Ithaca, N.Y., 1969.

[5] ———, *Continuously discounted Markov decision model with countable state and action space*, Ann. Math. Statist., to appear.

[6] B. L. MILLER, *Finite state continuous time Markov decision processes with an infinite planning horizon*, J. Math. Anal. Appl., 22 (1968), pp. 552–569.

[7] SHELDON M. ROSS, *Non-discounted denumerable Markovian decision models*, Ann. Math. Statist., 39 (1968), pp. 412–423.

[8] H. M. TAYLOR, *Markovian sequential replacement processes*, Ibid., 36 (1965), pp. 1677–1694.

[9] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, N.J., 1946.

Preface

This volume is based on papers presented at the NSF Regional Conference on Control Theory held at the University of Maryland Baltimore County, August 23–27, 1971. The purpose of the Conference was to have a gathering of experts, active researchers, and graduate students in the United States working or interested in the broad area of control theory.

The proceedings consisted of: (a) five two-hour lectures by Professor J. L. Lions of Paris on the control theory of partial differential equations with applications in science and engineering; (b) nine one-hour invited lectures by: Professors A. V. Balakrishnan of the University of California at Los Angeles; Roger W. Brockett of Harvard University; Wendell Fleming of Brown University; Henry Hermes of the University of Colorado at Boulder; Sanjoy Mitter of M.I.T.; Lucien Neustadt of the University of Southern California and Brown University; and Stephen G. Jones, V. M. Popov, and Leonard Weiss of the University of Maryland; and (c) twelve twenty-minute presentations by various authors. The talks by the invited speakers covered collectively major areas of interest with emphasis on system theoretic and geometric aspects of control theory and the control theory of stochastic, hereditary, and discrete-time systems.

The present volume fairly represents the proceedings except for the lectures of Professor Lions which will appear as a separate publication of SIAM in the NSF Regional Conference Series on Applied Mathematics. Other exceptions are the lectures by Professors Fleming, Jones, and Neustadt and some short presentations including those by Raymond Rischel of Bell Laboratories and Prem Prakash of Northwestern University. These latter have all appeared or will appear elsewhere.

The present volume is offered with the hope that it will stimulate high quality research in various aspects of control theory and its applications, thus broadening and deepening the scope of modern control theory and its applications in various directions of human endeavor.

The Conference was supported by the National Science Foundation under Grant NSF GP-30307, and by the University of Maryland Baltimore County with funds received from the Martin Marietta Corporation. The organization committee consisted of Professors A. K. Aziz, R. C. Roberts, and N. P. Bhatia. Special thanks are due to Professor Lucien Neustadt for his whole-hearted support in the publication of this volume, the referees of the papers in this volume for their prompt and to the point reports, to our secretaries Mrs. Mildred J. Hern and Miss Kathleen M. Hadfield, the Organization Committee, the Martin Marietta Corporation and the National Science Foundation. They all singly and collectively made a dream come true.

University of Maryland Baltimore County
Baltimore, Maryland
November 1971

NAM P. BHATIA

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

THE GEOMETRY OF TIME-OPTIMAL CONTROL*

HENRY HERMES†

Abstract. Consider a linear control system $dx/dt = A(t)x + B(t)u$, $x(0) = x^0 \in \mathbb{R}^n$, and the problem of characterizing the boundary of the set of points attainable at some time t_1, which we denote $\partial\mathscr{A}(t_1)$. It is shown how this leads to a set of generalized Hamiltonian equations of the form $dx/dt \in R(t, x, p)$, $x(0) = x^0$; $dp/dt = -pA(t)$, $p(0) = \xi \in S^{n-1}$, the $(n-1)$-sphere. If $\varphi(\cdot, \xi)$, $\psi(\cdot, \xi)$ is a solution pair of these equations, we interpret $\varphi(t_1, \cdot)$ as a map of S^{n-1} into the collection of nonempty, compact, convex, subsets of $\partial\mathscr{A}(t_1)$. Then $\bigcup \{\varphi(t_1, \xi) : \xi \in S^{n-1}\} = \partial\mathscr{A}(t_1)$. The relation between the geometry of $\mathscr{A}(t_1)$ and whether the Hamiltonian equations are generalized or ordinary is discussed.

These results for the linear problem are used to motivate the generalized Hamiltonian equations associated with nonlinear control systems, together with geometric properties of their attainable sets.

Introduction. We shall consider a system with dynamics described by a set of n controlled, ordinary differential equations

$$\dot{x}(t) = f(t, x(t), u(t)), \qquad x(0) = x^0 \qquad (\dot{x}(t) = dx(t)/dt), \tag{1}$$

where the control function u may be selected from some given admissible set of controls, termed Ω. A solution of (1) corresponding to a choice of control $u \in \Omega$ will be denoted φ^u; its value at time t being $\varphi^u(t)$. For a typical time-optimal control problem, one is given a continuous (target) map $z\colon [0, \infty) \to E^n$ (E^n denoting Euclidean n space); the problem is to find a control $u^* \in \Omega$ such that $\varphi^{u^*}(t^*) = z(t^*)$ while for $t < t^*$, $z(t) \notin \{\varphi^u(t) : u \in \Omega\}$. If this is possible, u^* would be called the optimal control and t^* the optimal time.

We pause to examine the above formulation. Let $\mathscr{C}(E^k)$ denote the set of nonempty, compact, subsets of E^k endowed with the Hausdorff metric topology. A typical description of Ω is as follows. If $U : [0, \infty) \to \mathscr{C}(E^k)$ continuously, is given, Ω is the set of all Lebesgue measurable functions $u : [0, \infty) \to E^k$ such that $u(t) \in U(t)$, $t \geqq 0$. Now assume f is continuous in all arguments and define

$$R(t, x) = \{f(t, x, v) : v \in U(t)\}.$$

A lemma of Filippov [1] easily yields the following [2, p. 106]: If φ is an absolutely continuous function with $\dot{\varphi}(t) \in R(t, \varphi(t))$ p.p., there exists an admissible control $u \in \Omega$ such that $\dot{\varphi}(t) = f(t, \varphi(t), u(t))$ p.p. This shows that the exact nature of f and U is not fundamental to the problem, which we now give the following more general reformulation. Let $R : [0, \infty) \times E^n \to \mathscr{C}(E^n)$ be continuous. We replace (1) by the generalized differential equation

$$\dot{x}(t) \in R(t, x(t)), \qquad x(0) = x^0. \tag{2}$$

* Received by the editors August 27, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80302. This work was supported by the National Science Foundation under Grant GP 27957.

By a solution of (2) we shall mean an absolutely continuous function φ such that $\varphi(0) = x^0$ and $\dot{\varphi}(t) \in R(t, \varphi(t))$ p.p.

The distinguishing feature of the time-optimal control problem is that it depends only on the knowledge of the set of points attainable by all solutions at any given time. Specifically, let

$$\mathscr{A}_R(t) = \{\varphi(t) \in E^n : \varphi \text{ is a solution of (2)}\}.$$

Then the system can be controlled to reach the target z if and only if $z(t) \in \mathscr{A}_R(t)$ for some $t \geqq 0$, while an optimal time exists if and only if when $t^* = \inf\{t : z(t) \in \mathscr{A}_R(t)\}$, then $z(t^*) \in \mathscr{A}_R(t^*)$. Let $\operatorname{cl}\mathscr{A}_R(t)$ denote the closure of $\mathscr{A}_R(t)$. With mild assumptions on R in (2), $\operatorname{cl}\mathscr{A}_R(\cdot) : [0, \infty) \to \mathscr{C}(E^n)$ is continuous; thus, if $z(t_1)$ belongs to the interior of $\mathscr{A}_R(t_1)$, it easily follows that there is a $t < t_1$ with $z(t) \in \mathscr{A}_R(t)$. Thus the "geometry" of the boundary of $\mathscr{A}_R(t)$ plays an essential role in time-optimal control, and we shall concentrate on its properties.

One should probably enlarge the scope of what is usually called a time-optimal problem. For example, let $\mathscr{C}[0, T]$ denote the continuous functions from $[0, T]$ to E^n with the uniform topology, and define

$$\mathscr{B}_R[0, T] = \{\varphi \in \mathscr{C}[0, T] : \varphi \text{ is a solution of (2)}\}.$$

Let $h : [0, T] \times \mathscr{B}_R[0, T] \to E^k$ continuously, assume $h(t, \cdot)$ is linear on $\mathscr{C}[0, T]$, and let $\mathscr{H}_R(t) = \{h(t, \varphi) : \varphi \in \mathscr{B}_R[0, T]\}$. If $h(t, \varphi) = \varphi(t)$, i.e., the evaluation map at time t, $\mathscr{H}_R(t) = \mathscr{A}_R(t)$. One could equally well seek the smallest $t \geqq 0$ such that $z(t) \in \mathscr{H}_R(t)$. One could also replace the target function z by a set-valued function $Z : [0, \infty) \to \mathscr{C}(E^n)$ continuously, and the condition $z(t) \in \mathscr{A}_R(t)$ becomes $Z(t) \cap \mathscr{A}_R(t) \neq \varnothing$. These generalizations do not yield essential differences in the theory and will not be considered here.

One could also replace the evolution equation, either (1) or (2), by one with values in a Banach space. Then $\mathscr{A}_R(t)$ becomes a subset of the Banach space and the geometry becomes more difficult. Such generalizations are necessary when dealing with functional differential equations (see [3], [4]) or with partial differential equations.

1. The attainable set for linear systems. We shall summarize some results for the case when (1) is linear. Here the geometry is quite complete and serves as good motivation for the nonlinear case. Consider

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \qquad x(0) = x^0, \tag{3}$$

where A is an $n \times n$ matrix-valued function with summable components and B is $n \times r$ matrix-valued with, we assume, essentially bounded components. Let $U : [0, \infty) \to \mathscr{C}(E^r)$ be given, continuous; then an admissible control will be a measurable function u with $u(t) \in U(t)$ p.p. If $X(\cdot)$ denotes a fundamental solution of $\dot{x}(t) = A(t)x(t)$ with $X(0) = I$, and we define $R(t) = \{X^{-1}(t)B(t)v : v \in U(t)\}$, then (by use of Filippov's lemma) it follows that x is a solution of (3) if and only if

$$x(t) = X(t)y(t), \tag{4}$$

where y is a solution of

$$\dot{y}(t) \in R(t), \qquad y(0) = x^0. \tag{5}$$

Since for each t, $X(t)$ is a nonsingular matrix, (4) shows a simple relation holds between the set of points attainable at any time t by solutions of (3), and the attainable set at time t, denoted $\mathscr{A}_R(t)$, of (5). Thus we study the latter.

We note that, as defined, $R\colon [0, \infty) \to \mathscr{C}(E^n)$ is measurable (see [2, § 8]) and integrably bounded in the sense of [5]. If $\mathrm{co}R\colon [0, \infty) \to \mathscr{C}(E^n)$ denotes the function with values $\mathrm{co}R(t)$, the convex hull of $R(t)$, Aumann's extension of Lyapunov's theorem on the range of a vector measure [5] yields:

For each t, $\mathscr{A}_R(t) = \mathscr{A}_{\mathrm{co}R}(t)$, and each of these is a compact, convex, subset of E^n.

Remark. It is not true that $\mathscr{B}_R[0, T]$, as defined in the Introduction, satisfies $\mathscr{B}_R[0, T] = \mathscr{B}_{\mathrm{co}R}[0, T]$. Indeed $\mathscr{B}_{\mathrm{co}R}[0, T]$ will be compact, while one can show that $\mathscr{B}_R[0, T]$ is compact if and only if $R(t)$ is convex for almost all t. (This follows from a slight modification of Lemma 1.4 in [6].) Since $\mathscr{A}_R(\cdot)$ is clearly continuous, an easy argument [2, p. 50] shows that in this case:

If there exists a $t_1 \geqq 0$ such that $z(t_1) \in \mathscr{A}_R(t_1)$, then there exists an optimal time and hence an optimal control for (3).

On the other hand, since $\mathscr{B}_R[0, T]$ is not necessarily compact, if the problem were to minimize some continuous cost function $F\colon \mathscr{C}[0, T] \to E^1$ over the trajectories $\mathscr{B}_R[0, T]$ of the system (3) (or (5)), clearly F need not attain a minimum on $\mathscr{B}_R[0, T]$.

Returning to the time-optimal problem for (5), we may note that it suffices to consider R merely measurable, integrably bounded and compact-valued, i.e., we need not refer to (3) at all. Since $\mathscr{A}_R(t_1)$ is compact and convex, it will have a support hyperplane at each point q of its boundary, denoted $\partial\mathscr{A}_R(t_1)$. Let q be such a point, y a solution of (5) such that $y(t_1) = q$ and $\pi(y(t_1))$ a support plane to $\mathscr{A}_R(t_1)$ at $y(t_1)$ with outward normal η. Then the inner product

$$\eta \cdot \int_0^{t_1} (\dot{y}(\tau) - r(\tau))\, d\tau \geqq 0 \tag{6}$$

for any measurable selection r of R (i.e., r is measurable with $r(\tau) \in R(\tau)$ p.p.) from which it follows that $\eta \cdot \dot{y}(\tau) = \max\{\eta \cdot r : r \in R(\tau)\}$ p.p. In terms of (3), substituting for y from (4) and letting $p(t) = \eta X^{-1}(t)$, this merely gives the maximum principle which we state as follows:

A necessary condition that a solution φ^{u^*} of (3) *satisfy $\varphi^{u^*}(t_1)$ belongs to the boundary of the attainable set, at time t_1, of* (3), *is that u^* satisfies, p.p., $p(t)B(t)u^*(t) = \max\{p(t) \cdot B(t)v : v \in U(t)\}$ where p satisfies the equation*

$$\dot{p}(t) = -p(t)A(t). \tag{7}$$

We see, also, that if $\varphi^{u^*}(t_1)$ belongs to the boundary of the attainable set at time t_1, $\varphi^{u^*}(t)$ belongs to the boundary of the attainable set at time t for all $0 \leqq t \leqq t_1$. In summary, and form which carries to the nonlinear case, let $H(t, x, p, u) = p \cdot A(t)x + p \cdot B(t)u$ and $H^*(t, x, p) = \max\{H(t, x, p, u); u \in U(t)\}$. Let $\varphi(\cdot, \xi)$, $\psi(\cdot, \xi)$ denote a solution pair to the equations

$$\begin{aligned} \dot{x}(t) &= \frac{\partial H^*}{\partial p}(t, x, p), & x(0) &= x^0, \\ \dot{p}(t) &= -\frac{\partial H^*}{\partial x}(t, x, p), & p(0) &= \xi \in S^{n-1}, \end{aligned} \tag{8}$$

where S^{n-1} denotes the $(n-1)$-sphere. Then any point of the form $\varphi(t_1, \xi)$, $\xi \in S^{n-1}$, belongs to the boundary of the attainable set at time t_1. (We remark that this property depends on the convexity of the attainable set for linear problems and does not, in general, hold for nonlinear problems.)

Let $\mathscr{A}(t_1)$ denote the attainable set, at time t_1, for the linear system (3). It is of interest to examine the map

$$\varphi(t_1, \cdot): S^{n-1} \to \partial\mathscr{A}(t_1) \subset E^n. \tag{9}$$

We shall list its properties; the verifications use mainly inequality (6) and can be found (essentially) in [2].

P1(a) If $q \in \partial\mathscr{A}(t_1)$ is an exposed point of $\mathscr{A}(t_1)$, q is in the range of the map $\varphi(t_1, \cdot)$.

P1(b) If $q \in \partial\mathscr{A}(t_1)$ is an exposed point at which $\mathscr{A}(t_1)$ has a unique support hyperplane, there exists a unique $\xi^* \in S^{n-1}$ such that $\varphi(t_1, \xi^*) = q$.

Geometrically, if $\mathscr{A}(t_1)$ has a "corner" at q, there will be one value $\xi \in S^{n-1}$ corresponding to each support hyperplane of $\mathscr{A}(t_1)$ at q. Each such value of ξ will determine the same (and unique) control u^* for (3), such that $\varphi^{u^*}(t_1) = q$. (See [2, Thm. 15.1].)

If $q \in \partial\mathscr{A}(t_1)$ is such that a support hyperplane of $\mathscr{A}(t_1)$ at q intersects $\mathscr{A}(t_1)$ in more than one point, the intersection is a convex set (a flat). In this case the inequality (6), analogously the maximum principle, does not determine a unique value $u \in U(t)$ which maximizes $H(t, x, p, u)$. Specifically, there will be a $\xi \in S^{n-1}$ and sets of positive t-measure for which $\max\{\psi(t, \xi)B(t)v: v \in U(t)\}$ does not determine a unique value $u(t) \in U(t)$. It is precisely in this case that the necessary condition is "ambiguous," and one may well argue that the map $\varphi(\cdot, \xi)$ is not well-defined for such values of ξ. Indeed, one should proceed as follows. Define

$$R(t, x, p) = \{A(t)x + B(t)u^*: p \cdot B(t)u^* = \max_{u \in U(t)} p \cdot B(t)u\}$$

and replace equations (8) by

$$\begin{aligned} &\dot{x}(t) \in R(t, x, p), && x(0) = x^0, \\ &\dot{p}(t) = -p(t)A(t), && p(0) = \xi \in S^{n-1}. \end{aligned} \tag{10}$$

Remark 1. The first of equations (10) is "linear," i.e., of the form $\dot{x}(t) \in \{A(t)x + q: q \in Q(t)\}$ with $Q: [0, \infty) \to \mathscr{C}(E^n)$ a measurable set-valued function; hence there is no problem concerning existence of solutions, and its associated attainable set, at any fixed time, is compact and convex.

Remark 2. If $\mathscr{A}(t_1)$ is strictly convex, one may easily show that (8) and (10) are the same on the interval $[0, t_1]$.

The above remarks point to a proper interpretation of the map $\varphi(t_1, \cdot)$ of (9). Let $\mathscr{C}(\partial\mathscr{A}(t_1))$ denote the set of nonempty, compact (and convex in this case) subsets of $\partial\mathscr{A}(t_1)$. Then we consider $\varphi(\cdot, \xi), \psi(\cdot, \xi)$ *redefined* as a solution pair for the equations (10), where ψ is point-valued but $\varphi(t_1, \xi)$ is to be interpreted as the attainable set, at time t_1, of all solutions of the first of equations (10). In this setting,

$$\varphi(t_1, \cdot): S^{n-1} \to \mathscr{C}(\partial\mathscr{A}(t_1)). \tag{11}$$

Now if $q \in \partial\mathscr{A}(t_1)$ and belongs to a "flat," i.e., a support hyperplane π intersects $\mathscr{A}(t_1)$ in points other than q, there will be a $\xi^* \in S^{n-1}$ such that $\varphi(t_1, \xi^*)$ *is* the compact, convex, set $\mathscr{A}(t_1) \cap \pi$.

In the first two listed properties of the map $\varphi(t_1, \cdot)$, we dealt with exposed points of $\partial\mathscr{A}(t_1)$; hence, the interpretations (9) and (11) of $\varphi(t_1, \cdot)$ are equivalent in that we can consider $\varphi(t_1, \xi)$ as a set consisting of a single point. We proceed to list properties, but now using the extended interpretation (11).

P1(c) $\bigcup \{\varphi(t_1, \xi) : \xi \in S^{n-1}\} = \partial\mathscr{A}(t_1)$.

P1(d) If q is an exposed point of $\mathscr{A}(t_1)$ at which $\mathscr{A}(t_1)$ has a corner (nonunique support planes) and $\varphi(t_1, \xi) = q$, then $\varphi(t, \xi)$ is a corner of $\mathscr{A}(t)$ for all $0 < t \leqq t_1$. (See [7], or [8], for more detailed information of this type.)

Thus, as time progresses, the boundary of $\mathscr{A}(t)$ may become "smoother," but may never develop "new corners."

P1(e) If there is a support plane of $\mathscr{A}(t_1)$ which intersects it in a k-dimensional flat, then for any $t_2 > t_1$, $\mathscr{A}(t_2)$ will have a support plane whose intersection with $\mathscr{A}(t_2)$ will contain a k-dimensional flat. (It is possible that this latter intersection could become, say, a $(k+1)$-dimensional flat; i.e., a strictly convex attainable set could, with increasing time, develop "flat spots.")

The geometric properties of $\mathscr{A}(t)$ listed above are essential in determining uniqueness of optimal control and optimal trajectory, in the time-optimal problem. A detailed account of this may be found in [2, §§ 15, 16] together with computable conditions (such as normality) which insure strict convexity of $\mathscr{A}(t)$. Sufficiency conditions can also be derived from the knowledge of $\mathscr{A}(t)$; indeed, t^* is a local minimum time if and only if the target z satisfies $z(t^*) \in \partial\mathscr{A}(t^*)$ and there exists an $\varepsilon > 0$ such that $z(t) \cap \mathscr{A}(t) = \varnothing$ for $t^* - \varepsilon < t < t^*$.

2. The geometry of the attainable set for nonlinear systems. It is convenient to have the dynamics described by either a system of ordinary differential equations, such as (1), or a generalized differential equation, such as (2). One may always reformulate (1) as a generalized differential equation; however, it is not always possible to go from (2) to a system of the form of (1).

For an arbitrary equation of the form (2), with R merely continuous, the existence of even a single solution remains an open question. If R is Lipschitz continuous (in the x variable) or continuous and convex-valued, or if it admits a representation of the form of (1), solutions do exist and these are the cases we shall consider. (See [9], [10] or [11] for questions concerning existence of solutions.) When we attempt to compare properties of the attainable set $\mathscr{A}_R(t_1)$ of (2) with those of a linear problem, we immediately find the following fundamental differences. First, $\mathscr{A}_R(t)$ need not be closed. Indeed, an example is given in [12] to show that it is even possible to have $\mathscr{A}_R(t)$ actually open. However, it is well known that if R has convex values in $\mathscr{C}(E^n)$, then $\mathscr{A}_R(t)$ will be closed. The second fundamental difference is that $\mathscr{A}_R(t)$ certainly need not be convex. This removes the possibility of the use of support hyperplanes, as in the linear case, and forces the analysis to depend more heavily on equations of the form (8) or (10).

We first shall examine the consequences of the lack of convexity of $\mathscr{A}_R(t)$. In doing this we shall keep other problems to a minimum by assuming the values $R(t, x)$, in (2), to be as "nice as possible." Specifically, we note that had the values $R(t)$ in (5) been strictly convex, then the maximization of $H(t, x, p, u)$, required for the maximum principle, would always have produced a unique value $u(t) \in U(t)$. This would have removed the possibility of "flat spots" on $\mathscr{A}(t)$ and the equations (10) would be superfluous. Thus, at this stage, we assume the following properties for the system (2):

(i) The values $R(t, x)$ are nonempty, compact, strictly convex, and the Gauss map (which associates to each point on the boundary of $R(t, x)$ the unit outward normal to the support plane at that point) is uniquely defined and continuously differentiable in terms of local coordinates.

(ii) The differential of the Gauss map has maximal rank at each point of $\partial R(t, x)$. (This merely gives positive curvature to the boundary of $R(t, x)$ and is a technicality needed to insure the differentiability of a map, r^*, defined below. For details, see [13, pp. 413–414].)

(iii) Equation (2) admits a formulation in the form of (1), with f smooth (at least C^2).

Roughly speaking, these requirements allow us to use $R(t, x)$ to determine a Minkowski metric in the tangent space, at (t, x), to our manifold which is just $[0, \infty) \times E^n$, and the time-optimal problem is equivalent to the geodesic problem in the resulting Finsler space. (See [14].) With these assumptions we can proceed with the maximum principle formulation, i.e., define $H(t, x, p, u) = p \cdot f(t, x, u)$ and if $u^* \in U(t)$ is such that it maximizes $H(t, x, p, u)$, let $r^*(t, x, p) = f(t, x, u^*)$. Then $r^*(t, x, p) \in \partial R(t, x)$ and the assumption (ii) above yields the differentiability of r^*. We now form the equations (analogous to (8))

$$
\begin{aligned}
\dot{x} &= r^*(t, x, p), & x(0) &= x^0, \\
\dot{p} &= -p \cdot r_x^*(t, x, p), & p(0) &= \xi \in S^{n-1},
\end{aligned}
\tag{12}
$$

where r_x^* denotes $\partial r^*/\partial x$. Again, let $\varphi(\cdot, \xi)$, $\psi(\cdot, \xi)$ denote a solution pair of these equations. We now define

$$S(t_1) = \{\varphi(t_1, \xi) : \xi \in S^{n-1}\}.$$

One may show that the map $(\varphi(t_1, \cdot), \psi(t_1, \cdot)) : S^{n-1} \to E^{2n}$ is differentiable and is, in fact, an imbedding, for any $t_1 > 0$. Thus $S(t_1)$ may be viewed as the projection, to the first n coordinates, of an imbedded $(n-1)$-sphere in E^{2n}. Some properties of $S(t_1)$ will next be listed. For proofs see [14].

P2(a) For any $t_1 > 0$, $\partial \mathscr{A}_R(t_1) \subset S(t_1)$, while for t_1 sufficiently small, $\varphi(t_1, \cdot) : S^{n-1} \to E^n$ is an imbedding and $\partial \mathscr{A}_R(t_1) = S(t_1)$.

What happens here is that as t_1 increases, the attainable set may "wrap around and overlap itself"; see Example 1 of [14].

P2(b) For t_1 sufficiently small so that $\varphi(t_1, \cdot)$ is an imbedding, $\psi(t_1, \xi)$ is an outward normal to $\partial \mathscr{A}_R(t_1)$ at the point $\varphi(t_1, \xi)$ for each $\xi \in S^{n-1}$.

Some remarks are in order. The fact that $\varphi(t_1, \cdot)$ is an imbedding for sufficiently small t_1 is analogous to the statement (in the study of geodesics) that locally every extremal is minimizing. This depends on the "roundness" assumptions on $R(t, x)$;

indeed, if R is merely convex-valued, an example which yields an extremal which does not minimize time for any $t_1 > 0$ is given in [15, Ex. 2.2]. In keeping with the terminology of the classical calculus of variations, a point $\varphi(t_1, \xi)$ is called *conjugate* to the point x^0 along the extremal $\varphi(\cdot, \xi)$ if rank $\partial\varphi(t_1, \xi)/\partial\xi < n - 1$. Thus conjugate points occur when the map $\varphi(t_1, \cdot)$ ceases to be an immersion (i.e., there is an extremal which is no longer "locally" minimizing).

P2(c) If $S(t_1)$ is an immersed sphere (i.e., $\varphi(t_1, \cdot): S^{n-1} \to E^n$ is an immersion), the degree of its normal (or Gauss) map is one. See Fig. 1.

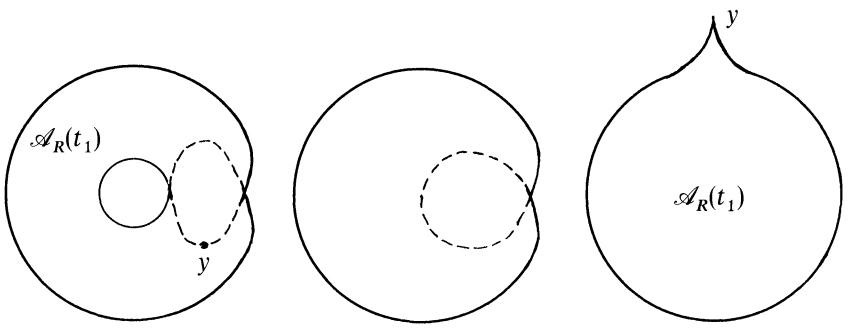


FIG. 1

Figure 1(a) can occur as the attainable set of a system of the form (2) (see [14, p. 61]); however, Fig. 1(b) cannot be an $S(t_1)$ since the degree of the Gauss map for the pictured immersed sphere would be two.

Can the figure indicated in Fig. 1(c) be the attainable set, at some time $t_1 > 0$, for a system of the form (2) satisfying the assumptions (i), (ii) and (iii) for R? The answer is no. Indeed, Roxin shows [7, p. 315] that in the nonlinear problem, as was the case in the linear problem, if the cone of normals to support hyperplanes to the attainable set $\mathscr{A}_R(t_1)$ at some boundary point $y = \varphi(t_1, \xi)$ consists of more than a single element, the same must be true at $\varphi(t, \xi)$ for $0 \leqq t \leqq t_1$. However, our "roundness" assumptions on the values of R imply, by property P2(a), that for some $0 < t' < t_1$, $\varphi(t', \cdot): S^{n-1} \to \partial\mathscr{A}_R(t')$ is a homeomorphism. From this and property P2(b), one can compose the inverse of this map with the homeomorphism $\psi(t', \cdot)/|\psi(t', \cdot)|: S^{n-1} \to S^{n-1}$ and obtain the fact that for t' sufficiently small, the Gauss map of $\partial\mathscr{A}_R(t')$ to S^{n-1} is a homeomorphism. This contradicts the property that the "cone of normals" at $\varphi(t', \xi)$ must consist of more than one element for some ξ; hence Fig. 1(c) cannot occur.

Returning to Fig. 1(a), we see that if $y \in S(t_1)$ but $y \notin \partial\mathscr{A}_R(t_1)$, then there is a solution $\varphi(\cdot, \xi)$ (an extremal) such that the maximum principle (12) holds, with $\varphi(t_1, \xi) = y$, yet y is an interior point of $\mathscr{A}_R(t_1)$. This demonstrates how the necessary condition can fail to be a sufficient condition. It is certainly conceivable that a figure such as Fig. 2 could occur.

In this case, $\mathscr{A}_R(t_1)$ would have a unique support hyperplane π at y, with $\pi \cap \mathscr{A}_R(t_1) = \{y\}$, yet there would not be a unique trajectory to y. (Compare this with property P1(b) for linear systems.)

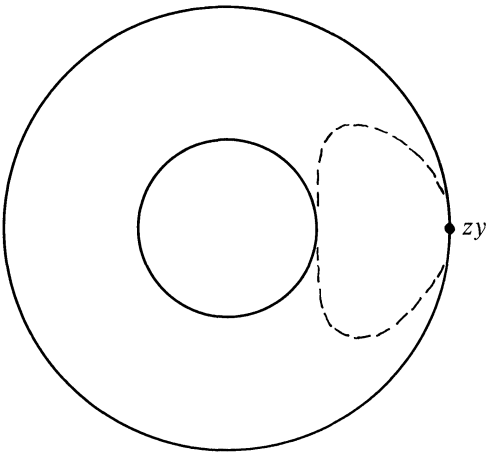


FIG. 2

Our next task is to examine what might occur if the strict conditions (i), (ii) and (iii) on R are relaxed. One might expect a new set of equations which generalize (12), in the same way that (10) generalized (8). However, one should recall that in the linear problem, say (5), the associated attainable set $\mathscr{A}_R(t_1)$ was always closed, even if R was not convex-valued. Thus, equations, such as (10), designed to produce solutions with values on $\partial\mathscr{A}_R(t)$, can be expected. In the nonlinear case, as previously noted, if the values $R(t, x)$ are not convex, $\mathscr{A}_R(t)$ need not be closed and in fact it may even be open. *In this case, no solution of* (2) *could produce a point on the topological boundary of* $\mathscr{A}_R(t)$. It is known [6] that if in (2) we replace R by the set-valued function co$R$ which has values co$R(t, x)$, then the relation between the associated attainable sets is $\mathscr{A}_{\text{co}R}(t) = \text{cl}\mathscr{A}_R(t)$. The conclusion is that if we expect to derive equations, similar to (12), whose solutions lead to points on the boundary of $\mathscr{A}_R(t)$, then we should at least require R to be convex-valued.

We now consider $R(t, x) = \{f(t, x, u) : u \in U(t)\}$ and assume the values $R(t, x)$ are convex, while $U : [0, \infty) \to \mathscr{C}(E^n)$ is continuous and f is smooth (say at least C^2). For any $p \in E^n - \{0\}$, let

$$H^*(t, x, p) = \max \{p \cdot f(t, x, u) : u \in U(t)\}$$

and

$$\tilde{R}(t, x, p) = \{f(t, x, u^*) : p \cdot f(t, x, u^*) = H^*(t, x, p)\}.$$

We now replace (12) with

$$\begin{aligned} &\dot{x} \in \tilde{R}(t, x, p), && x(0) = x^0, \\ &\dot{p} = -H_x^*(t, x, p), && p(0) = \xi \in S^{n-1}. \end{aligned} \tag{13}$$

As in the linear case, let $\varphi(\cdot, \xi)$, $\psi(\cdot, \xi)$ denote a solution pair and interpret $\varphi(t, \xi)$ as $\mathscr{A}_{\tilde{R}}(t)$. One may show that

$$\partial\mathscr{A}_R(t) \subset \bigcup_{\xi \in S^{n-1}} \varphi(t, \xi).$$

Very little work on equations of the form (13) has been done. These may be considered as "generalized Hamiltonian equations." Some related results for Lagrange problems with convex integrands have recently been studied by Rockafellar [16], [17].

REFERENCES

[1] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[2] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[3] H. T. BANKS AND M. Q. JACOBS, *The optimization of trajectories of linear functional differential equations*, this Journal, 8 (1970), pp. 461–488.

[4] H. T. BANKS, M. Q. JACOBS AND M. R. LATINA, *The synthesis of optimal controls for linear problems with retarded controls*, CDS Tech. Rep. 70–4, Brown Univ., Providence, 1970.

[5] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.

[6] H. HERMES, *The generalized differential equation* $\dot{x} \in R(t, x)$, Advances in Math., 4 (1970), pp. 149–169.

[7] E. ROXIN, *A geometric interpretation of Pontriagins maximum principle*, Proc. Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 303–324.

[8] H. HERMES, *Calculus of set valued functions and control*, J. Math. Mech., 18 (1968), pp. 47–60.

[9] A. F. FILIPPOV, *Classical solutions of differential equations with multivalued right-hand sides*, this Journal, 5 (1967), pp. 609–621.

[10] H. HERMES, *Existence and properties of solutions of* $\dot{x} \in R(t, x)$, Studies in Applied Math. 5, Soc. Indust. Appl. Math., Philadelphia, Pa., 1969, pp. 188–193.

[11] ———, *On continuous and measurable selections and the existence of solutions of generalized differential equations*, Proc. Amer. Math. Soc., 29 (1971), pp. 535–542.

[12] ———, *On the structure of attainable sets for generalized differential equations and control systems*, Differential Equations, 9 (1971), pp. 141–154.

[13] ———, *The equivalence and approximation of optimal control problems*, Ibid., 1 (1965), pp. 409–426.

[14] ———, *Attainable sets and generalized geodesic spheres*, Ibid., 3 (1967), pp. 256–270.

[15] ———, *Controllability and the singular problem*, this Journal, 2 (1965), pp. 241–260.

[16] R. T. ROCKAFELLAR, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.

[17] ———, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math., 33 (1970), pp. 411–427.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

CONTROLLABILITY, REALIZATION AND STABILITY OF DISCRETE-TIME SYSTEMS*

LEONARD WEISS†

Abstract. The following problems are discussed and solved in this paper: finding computable necessary and sufficient conditions for complete reachability and complete observability of a linear, time-varying, discrete-time system; finding sufficient conditions for local controllability of nonlinear discrete-time systems; relating reachability to the concept of discrete Pfaffian systems; obtaining a minimal-dimension difference equation (with possibly variable coefficients) from a given input/output function of a system; finding necessary and sufficient conditions for Lyapunov stability and finite-time stability of nonlinear difference equations; and, giving an algorithm for determining whether a linear difference equation is stable in the finite-time sense.

1. Introduction. In a certain sense, the theory of discrete-time systems dates back at least to the time of De Moivre and Laplace, who were the first to use the concept of generating functions in connection with the study of discrete random variables [1]. In the modern engineering literature, these functions are called Z-transforms [2], about which we shall say nothing further in this paper. The systematic study of difference equations (about which we shall say a good deal) began much later, with major landmarks in the development of the theory being provided by the treatises of Boole [3] and Milne-Thomson [4]. The interest in these equations in both mathematics and engineering stems from their usefulness in various applications. Numerical analysts deal with them in designing and analyzing algorithms for the numerical solution of differential and other equations [5], while engineers are often confronted with physical systems whose description by difference equations is quite natural [6]. In addition, any system whose internal structure may be unknown but whose input/output behavior can be (at least partially) obtained through experiment is a candidate for a difference equation model, and this accounts, in part, for the popularity of such models with economists, psychologists and statisticians (see [7] for many examples from these areas).

With the advent of digital computers, such models have become of compelling interest, and an accelerating growth of literature on various aspects of discrete-time systems has been the result [8].

In this paper, we study some selected problems in the mathematical theory of discrete-time systems, and we derive various results for both linear and nonlinear deterministic systems in the areas mentioned in the paper's title. (See [40] for a discussion of stochastic discrete-time systems.)

2. Preliminaries. Consider the difference equation

$$x(k+1) = f(k, x(k), u(k)), \qquad k \in \mathfrak{Z}, \tag{1}$$

* Received by the editors September 21, 1971, and in revised form October 28, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Department of Electrical Engineering and Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, College Park, Maryland 20742, and Mathematics Research Center, Naval Research Laboratory, Washington, D.C. This work was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-69-1646.

where $\mathfrak{Z}$ is the set of all integers, $x(k) \in \mathfrak{R}^n$, $u(k) \in \mathfrak{R}^p$, $p \leqq n$. Then, as long as f is a well-defined function on $\mathfrak{Z} \times \mathfrak{R} \times \mathfrak{R}^p$, there is no problem regarding existence and uniqueness of solutions to (1) starting from any given initial condition.

A linear discrete-time system is a system of the form (1) in which f is a linear function of x and u, i.e., (1) becomes

$$x(k+1) = A(k)x(k) + B(k)u(k), \qquad k \in \mathfrak{Z}, \tag{2}$$

where $A(k)$ is $n \times n$ and $B(k)$ is $n \times p$, $p \leqq n$.

In connection with (2), we define a real $n \times n$ matrix-valued function Ω on $\mathfrak{Z} \times \mathfrak{Z}$ by the formulas

$$\Omega(k,j) = A(k)A(k-1)\cdots A(j+1)A(j), \qquad j,k \in \mathfrak{Z}, \quad k \geqq j,$$

$$\Omega(k,k+1) \triangleq I \text{ (the identity)}, \qquad k \in \mathfrak{Z}, \tag{3}$$

$$\Omega(k,j) \quad \text{undefined for } j > k+1.$$

By iteration, the solution of (2) at the kth instant starting from initial time k_0 and state x_0 is

$$x(k; k_0, x_0, u) = \Omega(k-1, k_0)x_0 + \sum_{j=k_0}^{k-1} \Omega(k-1, j+1)B(j)u(j). \tag{4}$$

It should be noted that, unlike ordinary linear differential equations, it is possible for the set of all solutions of (2), with $u(\cdot) \equiv 0$, to be located in a proper subspace of $\mathfrak{R}^n$ (for example, take $A(\cdot) \equiv 0$). That is, discrete-time systems can be *pointwise degenerate* [9]. (If the set of all solutions to (2), with $u = 0$, spans $\mathfrak{R}^n$, the system (2) is said to be *pointwise complete* [10].) This property plays an important role in seeking necessary conditions for controllability, as will be shown later on. The possibility of pointwise degeneracy adds interest to the study of linear discrete-time systems, for it means that theories developed for linear ordinary continuous-time systems (where pointwise completeness always holds [11]) may not have 1–1 correspondence to their analogues in the discrete-time case. It is therefore of interest to develop results for discrete-time systems which are independent of the properties of pointwise completeness or degeneracy (and to point out circumstances under which considerations of these properties cannot be avoided).

Finally, we denote the set of positive integers by $\mathfrak{Z}^+$, and we define a discrete interval $[\alpha, \beta]$, where $\alpha, \beta \in \mathfrak{Z}, \alpha < \beta$, as the set of integers $\{\alpha, \alpha+1, \cdots, \beta-1, \beta\}$.

3. The concepts of controllability and reachability. One of the most fundamental contributions to the mathematical theory of systems over the past fifteen years has been the formulation and characterization of the property of controllability, which was first done by Kalman [12] for linear, time-invariant, discrete-time systems. Since then, much development of the theory of controllability (and its conceptual partner, the theory of reachability) has occurred for differential equations (see [13], [14]). In the sequel, we present some results for reachability and controllability of time-varying linear and nonlinear discrete-time systems using a variety of techniques.

In the following definition, the *phase space* $\mathfrak{P} = \mathfrak{R}^n \times \mathfrak{Z}$.

DEFINITION 1. (a) For the system (1), the phase $(x, v) \in \mathfrak{P}$ is *reachable* (or, *N-step reachable*) if there exists $N \in \mathfrak{Z}^+$, and a control sequence $\mathcal{U} = \{u(v - N + 1), u(v - N + 2), \cdots, u(v)\}$ such that the phase $(0, v - N)$ is transferred to (x, v) under the action of $\mathcal{U}$ (written $(0, v - N) \xrightarrow{\mathcal{U}} (x, v)$).

(b) If, for all $x \in \mathfrak{R}^n$, (x, v) is reachable (or N-step reachable), then the system (1) is *completely* (*N-step*) *reachable* at time v.

(c) Complete (N-step) reachability with no time designated implies that (1) is *completely* (*N-step*) *reachable* at all times.

DEFINITION 2. The phase (x, v) is *controllable* (or *N-step controllable*) if there exists $N \in \mathfrak{Z}^+$ and $\mathcal{U} = \{u(v), u(v + 1), \cdots, u(v + N - 1)\}$ such that $(x, v) \xrightarrow{\mathcal{U}} (0, v + N)$.

Definitions of complete (N-step) controllability follow in an entirely analogous fashion from Definition 1.

THEOREM 1. *A necessary and sufficient condition for* (2) *to be completely M-step reachable at time v is that*

$$\operatorname{rank}[B(v-1), \Omega(v-1, v-1)B(v-2), \cdots, \Omega(v-1, v-M+1)B(v-M)] = n. \tag{5}$$

Proof. Sufficiency. Let

$$\mathcal{R}_k(v-1) = [B(v-1), \Omega(v-1, v-1)B(v-2), \cdots, \Omega(v-1, v-k+1)B(v-k)] \tag{6}$$

and suppose rank $\mathcal{R}_M(v - 1) = n$. The solution to (2) at time v starting from the zero state at time $v - M$ is

$$x(v; v-M, 0, u) = \mathcal{R}_M(v-1) \begin{bmatrix} u(v-1) \\ u(v-2) \\ \vdots \\ u(v-M) \end{bmatrix}, \tag{7}$$

or, simplifying the notation,

$$x_M(v) = \mathcal{R}_M(v-1) U_M(v), \tag{8}$$

where

$$U_M(v) = \begin{bmatrix} u(v-1) \\ \vdots \\ u(v-M) \end{bmatrix}.$$

Now, define an n-vector $V_M(v)$ by the relation

$$U_M(v) = \mathcal{R}'_M(v-1) V_M(v), \tag{9}$$

where the "prime" indicates transpose. Then, from (8) and (9),

$$V_M(v) = [\mathscr{R}_M(v-1)\mathscr{R}'_M(v-1)]^{-1}x_M(v), \tag{10}$$

and so, we can solve for $V_M(v)$ and thus obtain, from (9), the appropriate sequence of controls needed to reach any given $x_M(v)$.

Necessity. Suppose rank $\mathscr{R}_M(v-1) < n$ but the system (2) is completely M-step reachable at time v. Then there exists a nonzero vector $\eta \in \mathfrak{R}^n$ such that $\eta'\mathscr{R}_M(v-1) = 0$. Hence, premultiplying both sides of (7) with η' yields $\eta' x(v; v-M, 0, u) = 0$ regardless of u. Since the system is completely M-step reachable at time v, choose $\{u(v-M), \cdots, u(v-1)\}$ such that $x(v; v-M, 0, u) = \eta$. Then $\eta'\eta = 0$ which contradicts the assumption that $\eta \neq 0$.

COROLLARY 1. *The system* (2) *is completely M-step reachable at time v if and only if the rows of* $\Omega(v-1, v-k+1)B(v-k)$, *considered as functions of k, are linearly independent over the discrete k-interval* $[1, M]$.

Proof. Let

$$\Theta(v-1, v-k) = \Omega(v-1, v-k+1)B(v-k).$$

If the rows of $\Theta(v-1, v-k)$ are linearly dependent on $[1, M]$, then there exists an n-vector $\eta \neq 0$ such that

$$\eta'\Theta(v-1, v-k) = 0 \quad \text{for all integers } k \in [1, M]. \tag{11}$$

But (11) implies $\eta'\mathscr{R}_M(v-1) = 0$, where $\mathscr{R}_M$ is given by (6). Hence rank $\mathscr{R}_M(v-1) < n$, and, by Theorem 1, the system is not completely reachable at time v. Reversing the argument proves the converse.

Remark 1. The criterion (5) is also a sufficient condition for complete (M-step) controllability of (2) at time $v-M$. It is *not*, however, a *necessary* condition for controllability as defined in Definition 2 (with $N = M$) unless $A(\cdot)$ is invertible on the discrete interval $[v-M+1, v-1]$ (the pointwise completeness condition for linear discrete-time systems).

Remark 2. The proof of Theorem 1 shows that complete M-step reachability at time v implies the ability to reach any fixed state at time v from *any* given state (not just the origin) at time $v-M$.

Remark 3. Notice that complete M-step reachability at time v implies complete N-step reachability at time v for all integers $N \geqq M$. This statement is false if reachability is replaced by controllability unless $A(\cdot)$ is invertible for all integers $\geqq v+M$.

Remark 4. In the time-invariant case, (5) reduces to the standard condition

$$\text{rank } [B, AB, \cdots, A^{M-1}B] = n \quad \text{for some } M \leqq n. \tag{12}$$

It therefore follows that in an n-dimensional time-invariant system, complete reachability (or controllability) implies complete M-step reachability (or controllability), with M any integer satisfying (12). The minimum possible value of M is $n-p+1$, where p is the number of control variables.

4. Controllability for nonlinear discrete-time systems. Using Theorem 1 plus a technique developed by Lee and Markus [15], one can obtain a result for controllability of systems of the form (1). In this case, it is expedient to assume that

f is a differentiable function of its arguments although only the values of f and its derivatives at discrete instants are of interest. (Hence this type of result is applicable to sample-data systems.)

Consider the system (1) with $f(k,0,0) \equiv 0$ and f differentiable in all its arguments.

DEFINITION 3. The system (1) is *locally controllable* at time v if there exists a neighborhood, $\mathcal{N}_0$, of the origin in $\mathfrak{R}^n$ such that for every state $x \in \mathcal{N}_0$, there exists $M \in \mathfrak{Z}^+$ and a sequence, $\{u(v), u(v+1), \cdots, u(v+M-1)\}$, such that $x(v+M; v, x, u) = 0$.

Now consider the system (2), where

$$A(k) = \frac{\partial f}{\partial x}(k,0,0), \qquad B(k) = \frac{\partial f}{\partial u}(k,0,0).$$

THEOREM 2. *The system* (1) *is locally controllable at time* v *if the system* (2), *with* $A(\cdot)$ *and* $B(\cdot)$ *as given above, is completely controllable at time* v.

Proof. For simplicity, let $v = 0$. For some fixed integer λ, and n-dimensional vector parameter ξ, let

$$u(k,\xi) = B'(k)[\Omega(\lambda - 1, k+1)]'\xi, \qquad k = 0, 1, \cdots, \lambda - 1.$$

Define

$$x(k+1,\xi) = f(k, x(k,\xi), u(k,\xi)), \qquad k = 0, 1, 2, \cdots,$$

with $x(0,\xi) = 0$, and let

$$J(k) = \frac{\partial x}{\partial \xi}(k,\xi)|_{\xi=0}.$$

Notice that $u(k,0) = 0$ for all k so that $x(k,0) = 0$ for all k. In that case, it is easy to show that J satisfies the difference equation

$$\begin{aligned} J(k+1) &= A(k)J(k) + B(k)\frac{\partial u}{\partial \xi}(k,\xi)|_{\xi=0} \\ &= A(k)J(k) + B(k)B'(k)[\Omega(\lambda - 1, k+1)]', \end{aligned} \tag{13}$$

where $J(0) = 0$ and $k = 0, 1, \cdots, \lambda - 1$. Iterating (13), we obtain

$$J(\lambda) = \mathscr{R}_\lambda(\lambda - 1)\mathscr{R}'_\lambda(\lambda - 1), \tag{14}$$

where $\mathscr{R}_\lambda(\lambda - 1)$ is given by (6). By hypothesis, there exists an integer M such that $\operatorname{rank}\mathscr{R}_M(M-1) = n$. Then $J(M)$ is nonsingular and the implicit function theorem allows one to obtain a solution to the equation $x(M; 0, x_0, \xi) = 0$ in terms of a mapping $\Pi: \mathcal{N}_0 \to \mathfrak{R}^n$ such that $\xi = \Pi(x_0)$. Hence, (1) is locally controllable at time 0.

5. An alternative approach to reachability. In this section we show that a discrete version of Pfaffian systems can be used to generate results on reachability for linear discrete-time systems.

Consider the system (2) and let $G(k)$ be an $(n-p) \times n$ matrix such that $\operatorname{rank} G(k) = n - \operatorname{rank} B(k)$ and $G(k)B(k) = 0$ for all integers $k \in [v - M, v - 1]$,

for some $M \in \mathfrak{Z}^+$, $M > 1$. Then, from (2), we have that

$$G(k)x(k+1) - G(k)A(k)x(k) = 0 \tag{15}$$

for all k on the discrete interval $[v - M, v - 1]$. We shall call (15) the *discrete Pfaffian* associated with (2). Let g_i' denote the ith row of G and let $g'(k)$ be an arbitrary nonzero linear combination of the rows of $G(k)$. That is,

$$g'(k) = \sum_i \alpha_i(k) g_i(k).$$

DEFINITION 4. The discrete Pfaffian system (15) is *summable* on $[v - M, v - 1]$ if and only if there exists some nonzero $g'(k)$ such that the expression

$$g'(k)x(k+1) - g'(k)A(k)x(k) \tag{16}$$

is an exact forward difference on $[v - M, v - 2]$.

More precisely, summability of (15) on $[v - M, v - 1]$ means that there exists a scalar-valued function $\varphi(k, x)$, where $(k, x) \in [v - M, v - 2] \times \mathfrak{R}^n$, such that

$$\begin{aligned} \Delta_x \varphi(k, x) &= g'(k), \\ \Delta_k \varphi(k, x) &= g'(k)[A(k) - I]x, \end{aligned} \tag{17}$$

where Δ_k represents the forward difference operator with respect to the variable k and

$$\Delta_x \varphi(k, x) = (\Delta_{x_1} \varphi(k, x), \cdots, \Delta_{x_n} \varphi(k, x)).$$

Clearly, this type of definition will also work for nonlinear systems of the form (1), provided the control $u(k)$ appears linearly. In the case at hand, the choice of φ is obvious, since, by inspection of (16), we have the following.

PROPOSITION 1. *A necessary and sufficient condition for* (16) *to be an exact forward difference on* $[v - M, v - 2]$ *is that*

$$g'(k-1) = g'(k)A(k)$$

for all k *on the discrete interval* $[v - M + 1, v - 1]$.

The function φ in (17) can therefore be taken as $\varphi(k, x) = g'(k)x$.

The main result we wish to prove in this section is as follows.

THEOREM 3. *The following statements are equivalent.*

(i) *The system* (2) *is completely* M*-step reachable at time* v.

(ii) *The discrete Pfaffian* (15) *is nonsummable on some discrete subinterval of* $[v - M, v - 1]$.

(iii) $\operatorname{Rank} \mathscr{R}_M(v - 1) = n$, *where* $\mathscr{R}_M$ *is given by* (16).

Proof. (i) $\Leftrightarrow$ (iii): This was proved as Theorem 1.

(i) $\Rightarrow$ (ii): Suppose the Pfaffian is summable on the entire discrete interval $[v - M, v - 1]$. Then there exists a nonzero row vector $g'(k)$ such that $g'(k)B(k) = 0$ for all integers $k \in [v - M, v - 1]$, and by Proposition 1, $g'(k-1) = g'(k)A(k)$

for all $k \in [\nu - M + 1, \nu - 1]$. Hence,

$$\begin{aligned} g'(\nu-1)\mathscr{R}_M(\nu-1) &= g'(\nu-1)[B(\nu-1), \Omega(\nu-1,\nu-1)B(\nu-2), \\ &\qquad \cdots, \Omega(\nu-1,\nu-M+1)B(\nu-M)] \\ &= [g'(\nu-1)B(\nu-1), g'(\nu-2)B(\nu-2), \cdots, g'(\nu-M)B(\nu-M)] \\ &= 0. \end{aligned}$$

But this implies that rank $\mathscr{R}_M(\nu-1) < n$, so that the system (2) is not completely M-step reachable at time ν. Taking the contrapositive establishes that (i) $\Rightarrow$ (ii).

(ii) $\Rightarrow$ (iii): Suppose rank $\mathscr{R}_M(\nu-1) = q < n$. Then there exists an $n \times n$ nonsingular matrix T such that

$$T\mathscr{R}_M(\nu-1) = \begin{bmatrix} \hat{\mathscr{R}}_M(\nu-1) \\ 0 \end{bmatrix}, \tag{18}$$

where $\mathscr{R}_M$ has q rows and rank $\hat{\mathscr{R}}_M(\nu-1) = q$. Let η' be a fixed nonzero row vector whose first q entries are 0. Define

$$g'(k) = \eta' T\Omega(\nu-1, k+1), \qquad k \in [\nu - M, \nu - 1]. \tag{19}$$

Then

$$g'(\nu-1) = \eta' T$$

and

$$g'(k-1) = g'(k)A(k), \qquad k \in [\nu - M + 1, \nu - 1]. \tag{20}$$

It then follows from (18) and (19) that $g'(k)B(k) = 0$ for all integers $k \in [\nu - M, \nu - 1]$, and (20) implies that the discrete Pfaffian associated with the system (2) is summable on the entire discrete interval $[\nu - M, \nu - 1]$. Taking the contrapositive proves (ii) $\Rightarrow$ (iii), which proves the theorem.

For a discussion of the Pfaffian technique applied to continuous-time systems, the reader is referred to Hermes [16] and Weiss [9].

6. Observability. The duality principle discovered by Kalman [12] is a statement of the fact that the mathematical structure of the optimal control, quadratic cost problem in control theory, and the optimal estimation, minimum variance problem in filtering theory, are identical for linear ordinary differential equations. The role of observability in filtering theory is completely dual to that of reachability in control theory.

The appropriate model for our study of observability in the discrete-time case is as follows:

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k), \\ y(k) &= C(k)x(k), \end{aligned} \tag{21}$$

where $k \in \mathfrak{Z}$, $y(k) \in \mathfrak{R}^m$ and represents the output of the system. $A(\cdot)$, $B(\cdot)$, $x(\cdot)$, $u(\cdot)$ are as in (2).

DEFINITION 5. (a) The system (21) is *completely* (N-*step*) *observable* at time μ if and only if there exists $N \in \mathfrak{Z}^+$ such that knowledge of $y(\mu), y(\mu+1), \cdots, y(\mu+N-1)$ and $u(\mu), u(\mu+1), \cdots, u(\mu+N-2)$ is sufficient to determine $x(\mu)$.

(b) The system (21) is *completely* (*N-step*) *determinable* at time μ if and only if there exists $N \in \mathfrak{Z}^+$ such that any state at time μ can be determined from knowledge of $y(\mu - N + 1), \cdots, y(\mu)$ and $u(\mu - N + 1), \cdots, u(\mu - 1)$.

(c) Complete observability (or determinability) without any time designation denotes *complete observability* (or *determinability*) at all times.

The definition of determinability differs from that of observability in that in the former case, we determine the "present" state from "past" measurements, while in the latter case, we determine a "past" state from "future" measurements.

THEOREM 4. *The system* (21) *is completely N-step observable at time* μ *if and only if*

$$\text{rank } [C'(\mu), [\Omega(\mu, \mu)]'C'(\mu + 1), \cdots, [\Omega(\mu + N - 2, \mu)]'C'(\mu + N - 1)] = n. \tag{22}$$

Proof. Sufficiency. The solution to (21) at the mth instant starting from initial time μ and initial state $x(\mu)$ is

$$y(m; \mu, x(\mu), u) = C(m)\Omega(m - 1, \mu)x(\mu) + \sum_{k=\mu}^{m-1} C(m)\Omega(m - 1, k + 1)B(k)u(k). \tag{23}$$

Let

$$\tilde{y}(m, \mu) = y(m; \mu, x(\mu), u) - \sum_{k=\mu}^{m-1} C(m)\Omega(m - 1, k + 1)B(k)u(k).$$

Then

$$\tilde{y}(m, \mu) = C(m)\Omega(m - 1, \mu)x(u), \qquad m = \mu, \mu + 1, \cdots. \tag{24}$$

Let

$$\mathscr{Y}_N(\mu) = \begin{bmatrix} \tilde{y}(\mu, \mu) \\ \tilde{y}(\mu + 1, \mu) \\ \vdots \\ \tilde{y}(\mu + N - 1, \mu) \end{bmatrix} \tag{25}$$

and let

$$\mathscr{O}_N(\mu) = [C'(\mu), [\Omega(\mu, \mu)]'C'(\mu + 1), \cdots, [\Omega(\mu + N - 2, \mu)]'C'(\mu + N - 1)]. \tag{26}$$

Then it follows from (24)–(26) that

$$x(\mu) = [\mathscr{O}_N(\mu)\mathscr{O}_N'(\mu)]^{-1}\mathscr{O}_N(\mu)\mathscr{Y}_N(\mu) \tag{27}$$

so that $x(\mu)$ can be computed as long as rank $\mathscr{O}_N(\mu) = n$.

Necessity. Suppose rank $\mathscr{O}_N(\mu) < n$ but the system (21) is completely N-step observable at time μ. Then there exists a nonzero vector $\xi \in \mathfrak{R}^n$ such that $\xi'\mathscr{O}_N(\mu) = 0$. From (24) and (25) we have

$$\mathscr{Y}_N(\mu) = \mathscr{O}_N'(\mu)x(\mu). \tag{28}$$

Setting $x(\mu) = \xi$ implies $\mathscr{Y}_N(\mu) = 0$ which violates complete N-step observability (the "output" is identically zero over $[\mu, \mu + N - 1]$ although the state at time μ is not zero).

Remark 5. The duality between Theorems 1 and 4 is obvious. The system (21) is completely N-step reachable (or completely N-step observable) at time v if and only if the system

$$\begin{aligned} z(k-1) &= A'(k)z(k) + C'(k)\tilde{u}(k), \\ \tilde{y}(k) &= B'(k)z(k), \end{aligned} \tag{29}$$

with the time scale reversed about v, is completely N-step observable (or completely N-step reachable) at time v.

Remark 6. One would expect the criterion for determinability (see Definition 5(c)) to be similar to that for observability (see (22)), but there is an important difference. The criterion (22) is only *sufficient* for complete N-step determinability at μ unless the matrix $A(\cdot)$ is nonsingular over $[\mu, \mu + N - 1]$. That is, pointwise degeneracy could force the "present" state to be zero regardless of "past" values of y. Since knowledge of the homogeneous system equation is presumed to be available to the observer, one could then, under the aforementioned circumstances, determine the "present" state regardless of the rank of the determinability matrix.

Remark 7. From condition (22) it is evident that complete N-step observability at time μ implies complete M-step observability at time μ for any integer $M \geqq N$. This is not the case for determinability, however, unless the pointwise completeness condition holds at all integral times $\leqq \mu - N$.

Remark 8. The discussion in Remarks 1, 5, 6 and 7 indicates that the pairing of reachability–observability and controllability–determinability as dual variables is most natural. This will become still clearer in the next section. For remarks on this problem within the continuous-time framework, where the issue is slightly less transparent, see Weiss [17] and Kalman [18].

Finally, the dual result of Corollary 1 is given.

COROLLARY 2. *The system* (21) *is completely N-step observable at time* μ *if and only if the columns of* $C(\mu + k)\Omega(\mu + k - 1, \mu)$, *considered as functions of* k, *are linearly independent over the discrete k-interval* $[0, N - 1]$.

7. Realization of input/output functions for linear discrete-time systems. As a result of different analysis or design considerations, dynamical systems are represented in various ways, e.g., by means of transfer functions, impulse responses and weighting patterns, input/output operator equations, state-variable equations, etc. For any kind of system with such different representations, it is desirable to be able to move easily from one representation to another. Exactly how one does this in the case of finite-dimensional systems has been the subject of many papers in recent years (see Kalman [19], Weiss and Kalman [20], and Youla [21]). A complete solution to the problem of constructing a state-variable differential or difference equation from input/output functions of "smooth" linear, time-invariant, finite-dimensional systems was given by B. L. Ho [22] via a now well-known algorithm.

Our objective in this section is to develop the background for a (partial) solution to the following problem: Given a graph of a matrix function of two discrete variables, $W(k, l)$, find (if possible) a linear, discrete-time system of the form of (21), having minimal state dimension, which generates the given data as the graph of the system's "unit pulse response" (see Definition 6 below).

We begin by considering the linear system (21) with its concomitant solution (23).

DEFINITION 6. The *unit pulse response matrix* for the system (21) is the $m \times p$ matrix function $W(k, l)$ given by

$$W(k,l) = \begin{cases} C(k)\Omega(k-1, l+1)B(l), & k > l, \\ 0, & k \leqq l. \end{cases} \tag{30}$$

Remark 9. The name "unit pulse response" is suggested by the fact that, in (21), if $u = \operatorname{col}(u_1, \cdots, u_p)$, $y = (y_1, \cdots, y_m)$, and $W = (w_{ij})$, then the ith column, W_i, of W can be expressed as

$$W_i(n, r) = y(n; r, 0, u),$$

where $u(k) = \operatorname{col}(0, \cdots, 0, \delta_{kr}, 0, \cdots, 0)$, and the Kronecker delta δ_{kr} appears in the ith entry.

Remark 10. In ordinary linear differential systems, the kernel matrix $W(t, \tau)$ (referred to as the "weighting pattern" [23] in system theory) is defined for all t, τ while the "causal impulse response" $W_C(t, \tau)$ is defined as

$$W_C(t,\tau) = \begin{cases} W(t,\tau), & t \geqq \tau, \\ 0, & t < \tau. \end{cases}$$

Except for the case where the system (21) is pointwise complete for all $k \in \mathfrak{Z}$ (i.e., the "A" matrix is invertible for all k), the unit pulse response $W(k, l)$ is not naturally well-defined for $l > k$, and is arbitrarily set to 0 in the latter region. In this sense, the unit pulse response and the causal impulse response are analogous.

PROPOSITION 2. *The unit pulse response* (30) *of a system* (21) *is invariant under coordinate transformations.*

Proof. Consider an arbitrary coordinate transformation $\hat{x}(k) = T(k)x(k)$. Then $\{A(\cdot), B(\cdot), C(\cdot)\}$ is transformed into $\{\hat{A}(\cdot), \hat{B}(\cdot), \hat{C}(\cdot)\}$ according to the relations

$$\begin{aligned} \hat{A}(k) &= T(k+1)A(k)T^{-1}(k), \\ \hat{B}(k) &= T(k+1)B(k), \\ \hat{C}(k) &= C(k)T^{-1}(k). \end{aligned}$$

Then

$$\begin{aligned} \hat{W}(k,l) &= \hat{C}(k)\tilde{\Omega}(k-1, l+1)\hat{B}(l) \\ &= C(k)T^{-1}(k)T(k)\Omega(k-1, l+1)T^{-1}(l+1)T(l+1)B(l) \\ &= C(k)\Omega(k-1, l+1)B(l) = W(k,l). \end{aligned}$$

Q.E.D.

Now consider (30) and, for some fixed integer λ, write it as

$$W(k,l) = C(k)\Omega(k-1, \lambda+1)\Omega(\lambda, l+1)B(l), \qquad l \leqq \lambda < k.$$

Let

$$\Psi(k,\lambda) \triangleq C(k)\Omega(k-1, \lambda+1), \qquad k > \lambda, \tag{31}$$

$$\Theta(\lambda, l) \triangleq \Omega(\lambda, l+1)B(l). \qquad l \leqq \lambda. \tag{32}$$

Then

$$W(k,l) = \Psi(k,\lambda)\Theta(\lambda,l), \qquad l \leqq \lambda < k. \tag{33}$$

Let

$$n_\theta(\lambda) \triangleq \text{number of rows of } \Theta(\lambda,\cdot) \text{ which are linearly independent on } (-\infty,\lambda],$$

$$n_\psi(\lambda) \triangleq \text{number of columns of } \Psi(\cdot,\lambda) \text{ which are linearly independent on } [\lambda+1,\infty),$$

$$n_0 \triangleq \max_\lambda \min\{n_\theta(\lambda), n_\psi(\lambda)\}. \tag{34}$$

(Note that n_0 is uniquely determined by (34).)

Define $\hat{\lambda}$ by the relation

$$n_0 = \min\{n_\theta(\hat{\lambda}), n_\psi(\hat{\lambda})\}. \tag{35}$$

DEFINITION 7. The unit pulse response $W(k,l)$ in (30) is *globally reduced* if and only if $n_0 = n_\theta(\hat{\lambda}) = n_\psi(\hat{\lambda})$.

LEMMA 1. *Every nonzero unit pulse response has a globally reduced form.*

Proof. Let $W(k,l)$ be a unit pulse response given by (30) and suppose it is not globally reduced. Assume, without loss of generality, that $n_\psi(\hat{\lambda}) < n_\theta(\hat{\lambda}) < n$ = number of rows (columns) of Ω. Then there exists an $n \times n$ nonsingular constant matrix T_1 such that

$$T_1\Theta(\hat{\lambda},\cdot) = \begin{bmatrix} \Theta_1(\hat{\lambda},\cdot) \\ 0 \end{bmatrix},$$

where the $n_\theta(\hat{\lambda})$ rows of Θ_1 are linearly independent on $(-\infty,\hat{\lambda}]$. Partitioning Ψ conformably with Θ_1 and multiplying Ψ on the right by T_1^{-1}, we get

$$\begin{aligned} W(k,l) &= [\Psi_{11}(k,\hat{\lambda}) \quad \Psi_{12}(k,\hat{\lambda})]\begin{bmatrix} \Theta_1(\hat{\lambda},l) \\ 0 \end{bmatrix} \\ &= \Psi_{11}(k,\hat{\lambda})\Theta_1(\hat{\lambda},l). \end{aligned}$$

Since the $n_\theta(\hat{\lambda})$ columns of Ψ_{11} are not linearly independent on $[\hat{\lambda}+1,\infty)$, there exists an $n_\theta(\hat{\lambda}) \times n_\theta(\hat{\lambda})$ nonsingular matrix T_2 such that

$$W(k,l) = \Psi_{11}(k,\hat{\lambda})T_2T_2^{-1}\Theta_1(\hat{\lambda},l),$$

where

$$\Psi_{11}(k,\hat{\lambda})T_2 = [\Psi(k,\hat{\lambda}) \quad 0]$$

and the $n_\psi(\hat{\lambda})$ columns of $\tilde{\Psi}$ are linearly independent over $[\hat{\lambda}+1,\infty)$. Writing

$$\tilde{\Theta}(\hat{\lambda},l) = T_2^{-1}\tilde{\Theta}_1(\hat{\lambda},l)$$

we then have that

$$W(k,l) = \tilde{\Psi}(k,\lambda)\tilde{\Theta}(\lambda,l) = \tilde{C}(k)\hat{\Omega}(k-1,l+1)\tilde{B}(l)$$

which is globally reduced with $\tilde{C}$ of dimension $m \times n_0$ and $\tilde{B}$ of dimension $n_0 \times p$.

DEFINITION 8. A *global realization* of a unit pulse response $W(k, l)$ is a linear finite-dimensional discrete-time system, defined on $\mathfrak{Z}$, whose unit pulse response coincides with $W(k, l)$. A realization on a smaller time set is a *local realization.*

DEFINITION 9. A *minimal realization* of a globally reduced unit pulse response is one which has the lowest state-space dimension of all global realizations.

Now consider the following lemma.

LEMMA 2. (a) *A minimal realization of a globally reduced unit pulse response is completely reachable and completely observable at some time v.*

(b) *Conversely, any realization which is completely reachable and completely observable at some time v is minimal.*

Proof. (a) Let $W(k, l)$ be given by (33). Suppose there is no time such that the realization $\{A(\cdot), B(\cdot), C(\cdot)\}$ is completely reachable and completely observable. Then, by Corollaries 1 and 2, and for any integer λ, either the columns of $\Psi(\cdot, \lambda)$ or the rows of $\Theta(\lambda, \cdot)$ are linearly dependent on the infinite discrete interval $[\lambda + 1, \infty)$ or $(-\infty, \lambda]$, respectively. But this contradicts the assumption that W is globally reduced.

(b) Suppose $\{A(\cdot), B(\cdot), C(\cdot)\}$ is a nonminimal completely reachable and completely observable realization of a unit pulse response. Let $\{\hat{A}(\cdot), \hat{B}(\cdot), \hat{C}(\cdot)\}$ be a minimal realization. Then

$$C(k)\Omega(k - 1, l + 1)B(l) = \hat{C}(k)\hat{\Omega}(k - 1, l + 1)\hat{B}(l)$$

and both expressions represent the same globally reduced unit pulse response. But $\dim \Omega(\cdot) > \dim \hat{\Omega}(\cdot)$ and this contradicts the uniqueness of n_0 in (34).

COROLLARY 3. *The dimension of any minimal realization of a unit pulse response* (30) *is the number* n_0 *in* (34).

DEFINITION 10. Two realizations of a given unit pulse response are *algebraically equivalent* if and only if they are related by a coordinate transformation.

Although a unit pulse response is invariant under coordinate transformations, and all minimal realizations have the same dimension, it does not follow that all minimal realizations are algebraically equivalent. This is due to the existence of "extra degrees of freedom" in the choice of $C(k)$ or $B(l)$ provided by the arbitrary setting of $W(k, l) = 0$ for $l \geqq k$. To illustrate this, consider the scalar systems $\{a(\cdot), b(\cdot), c(\cdot)\}$ and $\{a(\cdot), b(\cdot), \hat{c}(\cdot)\}$ in which

$$a \equiv 1, \qquad b(k) = \begin{cases} 1, & k = 3, 4, \\ 0, & \text{otherwise}, \end{cases}$$

$$c(k) = \begin{cases} 1, & k = 5, 6, \\ 0, & \text{otherwise}, \end{cases} \qquad \hat{c}(k) = \begin{cases} 1, & k = 1, 2, 5, 6, \\ 0, & \text{otherwise}. \end{cases}$$

Then the unit pulse response for both systems is

$$W(k, l) = \begin{cases} 1, & k = 5, 6, \quad l = 3, 4, \\ 0, & \text{otherwise}, \end{cases} \tag{36}$$

but the two systems, which are both minimal realizations of the unit pulse response (36), are not algebraically equivalent.

We now delineate a class of systems of the form (21) in which all minimal realizations of the associated unit pulse responses are algebraically equivalent.

The proof of the following theorem is analogous to one given by Youla [21] (see also Kalman [13]) for a result on algebraic equivalence of continuous-time systems announced in [23] (see also [20]).

THEOREM 5. *Two realizations of a given unit pulse response are algebraically equivalent if they are completely N-step reachable and completely M-step observable for some* $N, M \in \mathfrak{Z}^+$.

Proof. Let $\{A_i(\cdot), B_i(\cdot), C_i(\cdot)\}$, $i = 1, 2$, be the two realizations. Let

$$\Psi_i(k, \lambda) = C_i(k)\Omega_i(k - 1, \lambda + 1), \qquad i = 1, 2,$$

$$\Theta_i(\lambda, l) = \Omega_i(\lambda, l + 1)B_i(l), \qquad i = 1, 2.$$

Then, for any $\lambda \in \mathfrak{Z}$, the rows of $\Theta_i(\lambda, \cdot)$ and the columns of $\Psi_i(\cdot, \lambda)$ are linearly independent on the discrete intervals $I_N = [\lambda - N + 1, \lambda]$ and $J_M = [\lambda + 1, \lambda + M - 1]$, respectively. Let

$$K_i = \sum_{k \in J_M} \Psi_i'(k, \lambda)\Psi_i(k, \lambda), \qquad i = 1, 2,$$

$$L_i = \sum_{k \in I_N} \Theta_i(\lambda, k)\Theta_i'(\lambda, k), \qquad i = 1, 2.$$

Then, since

$$W(k, l) = \Psi_i(k, \lambda)\Theta_i(\lambda, l), \qquad l \leqq \lambda < k, \quad i = 1, 2,$$

we have

$$\begin{aligned} \Theta_2(\lambda, l) &= K_2^{-1}\left[\sum_{k \in J_M} \Psi_2'(k, \lambda)\Psi_1(k, \lambda)\right]\Theta_1(\lambda, l) \\ &= U\Theta_1(\lambda, l), \end{aligned} \tag{37}$$

where

$$U = K_2^{-1} \sum_{k \in J_M} \Psi_2'(k, \lambda)\Psi_1(k, \lambda).$$

Similarly,

$$\begin{aligned} \Psi_2(k, \lambda) &= \Psi_1(k, \lambda)\left[\sum_{l \in I_N} \Theta_1(\lambda, l)\Theta_2'(\lambda, l)\right]L_2^{-1} \\ &= \Psi_1(k, \lambda)V, \end{aligned} \tag{38}$$

where

$$V = \left[\sum_{l \in I_N} \Theta_1(\lambda, l)\Theta_2'(\lambda, l)\right]L_2^{-1}.$$

Then

$$\Psi_1(k, \lambda)\Theta_1(\lambda, l) = \Psi_2(k, \lambda)\Theta_2(\lambda, l) = \Psi_1(k, \lambda)VU\Theta_1(\lambda, l).$$

Hence $VU = I$ or $U = V^{-1}$. It then follows from (37), (38), and the definitions of Θ_i and Ψ_i, that $\{A_1(\cdot), B_1(\cdot), C_1(\cdot)\}$ is algebraically equivalent to $\{A_2(\cdot), B_2(\cdot), C_2(\cdot)\}$.

8. Construction of minimal realizations. We now present an algorithm for constructing a minimal realization of a unit pulse response given in the form of numerical data. The algorithm, which will work whenever the given unit pulse response possesses a realization which is completely N-step reachable and completely N-step observable for some $N \in \mathfrak{Z}^+$, is structurally analogous to one given by Silverman [24] for certain special classes of continuous-time systems. Important computational advantages are gained in the present context, however, and the result can be viewed as part of a procedure for synthesizing optimal linear digital filters [25].

We begin by defining the generalized Hankel matrix (see [26]) for the unit pulse response $W(k, l)$ given by (30).

DEFINITION 11. The *generalized Hankel matrix* of a unit pulse response $W(k, l)$ is given by

$$\mathscr{W}_N(k, l) =$$

(39)
$$\begin{bmatrix} W(k,l) & W(k,l-1) & \cdots & W(k,l-N+1) \\ W(k+1,l) & W(k+1,l-1) & \cdots & W(k+1,l-N+1) \\ \vdots & & & \\ W(k+N-1,l) & W(k+N-1,l-1) & \cdots & W(k+N-1,l-N+1) \end{bmatrix}.$$

From (30), (6), and (26), we have

$$\mathscr{W}_N(k, k-1) = \mathscr{O}'_N(k)\mathscr{R}_N(k-1). \tag{40}$$

THEOREM 6. *Let $W(k, l)$ be an $m \times p$ matrix function of two discrete variables. Suppose W is a unit pulse response with an n-dimensional realization $\{A(\cdot), B(\cdot), C(\cdot)\}$ as in* (30), *and, for some $N \in \mathfrak{Z}^+$, $\mathscr{W}_N(k, k-1)$ contains a fixed $n \times n$ submatrix $\Gamma(k)$ which is nonsingular for all $k \in \mathfrak{Z}$. Then there exists an n-dimensional, completely N-step reachable and completely N-step observable realization of W, given by*

$$\{\Gamma_1\Gamma^{-1}, \Lambda, \Phi\Gamma^{-1}\}, \tag{41}$$

where

$\Lambda(k)$ = *submatrix of $\mathscr{W}_N(k, k-1)$ consisting of those rows of the first m-column block whose indices match those of the rows of Γ;*

$\Phi(k)$ = *submatrix of $\mathscr{W}_N(k, k-1)$ consisting of those columns of the first p-row block whose indices match those of the columns of Γ;*

$\Gamma_1(k)$ = *submatrix of $\mathscr{W}_N(k+1, k-1)$ consisting of those elements whose indices match those of the elements of Γ.*

Proof. Since we postulate existence of an n-dimensional realization and rank $\mathscr{W}_N(k, k-1) \geqq n$ for all k, it follows from (40) that rank $\mathscr{W}_N(k, k-1)$ = rank $\mathscr{O}_N(k)$ = rank $\mathscr{R}_N(k-1) = n$ for all k. But, by Theorems 1 and 2, this implies that the realization $\{A(\cdot), B(\cdot), C(\cdot)\}$ is completely N-step reachable and completely N-step observable, and by Lemma 2, this realization is minimal. Now, it follows from (40) that

$$\Gamma(k) = \Gamma_o(k)\Gamma_r(k-1), \tag{42}$$

where $\Gamma_o(k)$ consists of those rows of $\mathscr{O}'_N(k)$ whose indices match those of the rows of Γ, and $\Gamma_r(k-1)$ consists of those columns of $\mathscr{R}_N(k-1)$ whose indices match those of the columns of Γ. Since Γ is nonsingular, it follows that Γ_o and Γ_r are nonsingular. Defining $\Gamma_1(k)$ as the matrix contained in

$$\mathscr{W}_N(k+1, k-1) = \begin{bmatrix} C(k+1)A(k) \\ C(k+2)C(k+1)A(k) \\ \vdots \\ C(k+N-1)\cdots C(k+1)A(k) \end{bmatrix} [B(k-1), A(k-1)B(k-2), \cdots, A(k-1)\cdots A(k+N-1)B(k-N)],$$

the indices of whose elements match those of the elements of Γ, it is clear that

$$\Gamma_1(k) = \Gamma_o(k+1)A(k)\Gamma_r(k-1).$$

But since $\Gamma_1(k)$ is a submatrix of $\mathscr{W}_N(k, k-1)$ and rank $\mathscr{W}_N(k, k-1) = \text{rank } \mathscr{W}_{N+1} \cdot (k, k-1)$ it follows that there exists an $n \times n$ matrix $\hat{A}(k)$ such that

$$\Gamma_1(k) = \hat{A}(k)\Gamma(k). \tag{43}$$

Then

$$\Gamma_o(k+1)A(k)\Gamma_r(k-1) = \hat{A}(k)\Gamma_o(k)\Gamma_r(k-1)$$

which yields

$$\hat{A}(k) = \Gamma_o(k+1)A(k)\Gamma_o^{-1}(k) = \Gamma_1(k)\Gamma^{-1}(k). \tag{44}$$

In like manner, it follows from (40) plus the definitions of Λ and Φ that

$$\Lambda(k) = \Gamma_o(k+1)B(k) \tag{45}$$

and

$$\Phi(k) = C(k)\Gamma_r(k-1) = C(k)\Gamma_o^{-1}(k)\Gamma(k). \tag{46}$$

It is now a simple matter to note that $\{A(\cdot), B(\cdot), C(\cdot)\}$ is algebraically equivalent to $\{\Gamma_1\Gamma^{-1}, \Lambda, \Phi\Gamma^{-1}\}$, where the associated coordinate transformation matrix is Γ_o. The realization (41) is completely N-step reachable and completely N-step observable (since these properties are invariant under algebraic equivalence), and the realization is obviously minimal.

Remark 11. Although the algorithm proceeds from the assumption that a realization with the appropriate properties exists, one need not know what this realization is in order to obtain (41).

Remark 12. The realization (41) will be time-invariant if a time-invariant realization exists, since $\mathscr{W}_N(k, k-1)$ will be a constant matrix under those circumstances.

Remark 13. The algorithm will yield a realization defined on any discrete k-interval on which the function $\mathscr{W}_N(k, k-1)$ is defined.

Remark 14. The "partial realization" problem discussed by Kalman [27] and Tether [28] has an analogue in the present framework, but we shall not consider that problem here.

9. Lyapunov stability for discrete-time systems. Two of the most fundamental questions arising in control system analysis and design are:

(a) What are the possibilities for identification and alteration of system behavior?

(b) Is the system stable (in some sense)?

Up to this point, we have been concerned only with question (a) and its ramifications with respect to the realization problem. We now turn to question (b) and discuss some aspects of the stability problem for discrete-time systems.

We first consider the nonlinear, homogeneous, n-dimensional system

$$x(k+1) = f(k, x(k)), \qquad k \in \mathfrak{S} = \{k_0, k_0+1, k_0+2, \cdots\}, \tag{47}$$

defined within a region $\mathfrak{J} = \{x \in R^n : \|x\| \leqq H\}$. It is assumed that f is finite for all finite values of its arguments and that $f(k, 0, 0) = 0$ for all $k \in \mathfrak{S}$.

DEFINITION 12. The zero solution of the system (47) is *stable* if, for any $\varepsilon > 0$, there exists $\delta(\varepsilon, k_0) > 0$ such that $\|x(k_0)\| < \delta$ implies $\|x(k; k_0, x(k_0))\| < \varepsilon$ for all $k \in \mathfrak{S}$.

Sufficient conditions for stability of (47) in terms of existence of Lyapunov functions have been known for some time (see [29]). The converse problem, however, has been treated only sparsely in the literature, and usually only with very restrictive assumptions. For example, Hahn [30] has proved a converse theorem within the context of "general motions" of dynamical systems, but his assumptions, when applied to (47), essentially include the requirement that f be "invertible" (i.e., that $x(k)$ can be expressed as a function of $x(k+1)$).

We now show that this assumption is unnecessary for the concrete model (47) by proving the following Lyapunov-type stability theorem and its converse.

THEOREM 7 (Weiss and Lam [31]). *The system* (47) *is stable if and only if there exists a real-valued function* $V(k, x)$ *defined on* $(k, x) \in \mathfrak{S} \times \mathfrak{J}$ *and continuous in* x *at* $x = 0$, *such that*:

(i) $V(k, 0) = 0$ *for all* $k \in \mathfrak{S}$.

(ii) *There exists a real-valued, continuous, monotone-increasing positive definite function* $a(\cdot)$ *such that* $V(k, x) \geqq a(\|x\|)$ *for all* $k \in \mathfrak{S}$, *all* $x \in \mathfrak{J}$.

(iii) $\Delta V(k; x(k; k_0, x)) \leqq 0$ *for all* $k \in \mathfrak{S}$, $x \in \mathfrak{J}$, *where* Δ *is the forward difference operator.*

Proof. Sufficiency. By the assumptions on V, corresponding to any given $\varepsilon > 0$, $(\varepsilon < H)$, we can choose $\delta > 0$ such that $\|x_0\| < \delta$ implies $\|x(k_1; k_0, x_0)\| = \varepsilon' \geqq \varepsilon$. By hypothesis (iii),

$$V(k_1, x(k; k_0, x_0)) \leqq V(k_0, x_0).$$

Hence, we have

$$a(\varepsilon) \leqq a(\varepsilon') \leqq V(k_1, x(k_1; k_0, x_0)) \leqq V(k_0, x_0) < a(\varepsilon)$$

which is a contradiction.

Necessity. Let

$$V(k, x) = \min\{H; \sup_{\substack{j \geqq k \\ j,k \in \mathfrak{S}}} \{\|x(j; k, x)\|\}\},$$

$$a(\|x\|) = \|x\|.$$

Then $V(k,0) = 0$ for all $k \in \mathfrak{S}$, and since (47) is assumed to be stable V is continuous in x at $x = 0$. Moreover, it is simple to check that hypothesis (ii) is satisfied. Now, for any $k \in \mathfrak{S}$ and any $x \in \mathfrak{J}$, we have

$$\Delta V(k, x(k;k_0,x)) = \min\{H; \sup_{\substack{j \geq k+1 \\ j,k+1 \in \mathfrak{S}}} \{\|x(j;k+1,x(k+1;k_0,x))\|\}\}$$
$$- \min\{H; \sup_{\substack{j \geq k \\ j,k \in \mathfrak{S}}} \{\|x(j;k,x(k;k_0,x))\|\}\}.$$

Since

$$\sup_{\substack{j \geq k+1 \\ j,k+1 \in \mathfrak{S}}} \|\cdot\| \leqq \sup_{\substack{j \geq k \\ j,k \in \mathfrak{S}}} \|\cdot\|$$

we have that $\Delta V(k; x(k;k_0,x)) \leqq 0$ for all $k \in \mathfrak{S}$, all $x \in \mathfrak{J}$.

It should be noted that without further information about the behavior of trajectories of (47), the "sup" function alone cannot suffice as an appropriate V-function in order to prove the necessity of the hypotheses of Theorem 7. If one is interested in boundedness rather than stability, however, such a V-function is admissible [31].

10. Finite-time stability of nonlinear discrete-time systems. In certain practical situations, stability in the sense of Definition 12 may be irrelevant. It may be more pertinent to require some function of the state to remain bounded in a particular way over a fixed finite number of discrete instants rather than to consider asymptotic properties as $k \to \infty$. The theory of finite-time stability has been created in response to such types of problems within a continuous-time context.

We now develop some results in this theory (see Weiss and Lam [27]) for nonlinear discrete-time systems.

Let $\mathfrak{S}_N = \{k_0, k_0+1, \cdots, k_0+N\}$.

DEFINITION 13. The system (47) is *stable with respect to* $(\alpha, \beta, \mathfrak{S}_N)$, $\alpha < \beta$, if $\|x_0\| \leqq \alpha$ implies $\|x(k;k_0,x_0)\| < \beta$ for all $k \in \mathfrak{S}_N$.

DEFINITION 14. The system (47) is *uniformly stable with respect to* $(\alpha, \beta, \mathfrak{S}_N)$, $\alpha < \beta$, if $\|x\| \leqq \alpha$ implies $\|x(j;k,x)\| < \beta$ for $j \in \{k, k+1, \cdots, k_0+N\}$, for all $k \in \mathfrak{S}_N$.

DEFINITION 15. The system (47) is *contractively stable with respect to* $(\alpha, \beta, \gamma, \mathfrak{S}_N)$, $\beta < \alpha < \gamma$, if it is stable with respect to $(\alpha, \gamma, \mathfrak{S}_N)$ and $\|x_0\| \leqq \alpha$ implies $\|x(k_0+N;k_0,x_0)\| < \beta$.

We use the following notation:

$$V_m^a(k) = \min_{\|x\|=a} V(k,x), \qquad V_{\underline{m}}^a(k) = \min_{\|x\| \geqq a} V(k,x),$$
$$V_M^a(k) = \max_{\|x\|=a} V(k,x), \qquad V_{\overline{M}}^a(k) = \max_{\|x\| \leqq a} V(k,x),$$
$$\mathfrak{B}(a) = \{x \in R^n : \|x\| < a\}, \qquad \overline{\mathfrak{B}}(a) = \{x \in R^n : \|x\| \leqq a\}.$$

THEOREM 8. *The system* (47) *is stable with respect to* $(\alpha, \beta, \mathfrak{S}_N)$, $\alpha < \beta$, *if there exists a real-valued function* $V(k,x)$, *defined for all* $k \in \mathfrak{S}_N$, *and continuous in* $x \in \mathfrak{R}^n$, *and a real-valued function* $\varphi(k)$ *defined on* $\mathfrak{S}_{N-1}$, *such that*

(i) $\Delta V(k, x(k;k_0,x_0)) \leqq \varphi(k)$ *for all* $k \in \mathfrak{S}_{N-1}$, *all* $x \in \overline{\mathfrak{B}}(\beta)$,

(ii) $\sum_{\substack{j \in \mathfrak{S}_{N-1} \\ j<k}} \varphi(j) < V_{\underline{m}}^{\beta}(k) - V_{\overline{M}}^{\alpha}(k_0)$ *for all* $k \in \mathfrak{S}_N$.

Furthermore, if the function f in (47) *is "invertible", then the converse holds.*

Proof. It follows from (i) that for all $k \in \mathfrak{S}_N$,

$$V(k, x(k; k_0, x_0)) \leqq V(k_0, x_0) + \sum_{\substack{j \in \mathfrak{S}_{N-1} \\ j < k}} \varphi(j).$$

Suppose $\|x_0\| \leqq \alpha$ and there exists $l \in \mathfrak{S}_N$ (the first such point) such that $\|x(l; k_0, x_0)\| \geqq \beta$, while $\|x(l-1; k_0, x_0)\| < \beta$. Then

$$\begin{aligned} V_{\underline{m}}^{\beta}(l) \leqq V(l; x(l; k_0, x_0)) &\leqq V(k_0, x_0) + \sum_{\substack{j \in \mathfrak{S}_{N-1} \\ j < l}} \varphi(j) \\ &\leqq V_{\bar{M}}^{\alpha}(k_0) + \sum_{\substack{j \in \mathfrak{S}_{N-1} \\ j < l}} \varphi(j) \end{aligned}$$

which contradicts (ii). This proves the first part.

To prove the converse under the additional hypothesis on f, we take $\varphi(j) \equiv 0$ and define

$$V(k, x) = \|x(k_0; k, x)\|, \qquad k \in \mathfrak{S}_N, \quad x \in \mathfrak{R}^n.$$

Then V is continuous in x and

$$V(k, x(k; k_0, x_0)) = \|x(k_0; k, x(k; k_0, x_0))\| = \|x_0\|$$

for all $k \in \mathfrak{S}_N$.

Hence,

$$\Delta V(k, x) = 0 \quad \text{for all } k \in \mathfrak{S}_N, \quad \text{all } x \in \mathfrak{R}^n,$$

so (i) is satisfied. Now,

$$V(k_0, x) = \|x(k_0; k_0, x)\| = \|x\|$$

so that

$$V_{\bar{M}}^{\alpha}(k_0) = \alpha. \tag{48}$$

Since the system is stable with respect to $(\alpha, \beta, \mathfrak{S}_N)$, it follows that for any pair (k_1, x_1), with $k_1 \in \mathfrak{S}_N$, $\|x_1\| \geqq \beta$, we have

$$V(k_1, x_1) = \|x(k_0; k_1, x_1)\| > \alpha.$$

By continuity of V in x,

$$V_{\underline{m}}^{\beta}(k) > \alpha \quad \text{for all } k \in \mathfrak{S}_N. \tag{49}$$

Combining (48) and (49) yields condition (ii), and the theorem is proved.

For application of this theorem to finite-time stability of *linear* systems, see Weiss and Lee [32].

We now state, without proof, the corresponding theorem for uniform stability.

THEOREM 9. *The system* (47) *is uniformly stable with respect to* $(\alpha, \beta, \mathfrak{S}_N)$, $\alpha < \beta$, *if there exists a real-valued function* $V(k, x)$, *defined for all* $k \in \mathfrak{S}_N$, *and continuous in* $x \in \mathfrak{R}^n$, *and a real-valued function* $\phi(k)$ *defined on* $\mathfrak{S}_{N-1}$, *such that*:

(i) $\Delta V(k, x) \leqq \phi(k)$ *for all* $k \in \mathfrak{S}_{N-1}$, *all* $x \in \bar{\mathfrak{B}}(\beta)$,

(ii) $\sum_{j=k_1}^{k_2-1} \phi(j) < V_{\underline{m}}^{\beta}(k_2) - V_{\bar{M}}^{\alpha}(k_1)$ *for all* $k_1, k_2 \in \mathfrak{S}_N, k_2 > k_1$.

Furthermore, if f in (47) *is "invertible," the converse holds.*

Finally, we present a theorem on contractive stability for systems of the form (47). The result is analogous to that of Kayande [33] for differential equations, and is in the spirit of a result given by Hurt [34]. The latter gives some interesting applications of results of this type to error analysis in numerical computation. Since the proof requires only one extra step beyond that of Theorem 8, it is omitted.

THEOREM 10. *The system* (47) *is contractively stable with respect to* $(\alpha, \beta, \gamma, \mathfrak{S}_N)$, $\beta < \alpha < \gamma$, *if there exist a real-valued function* $V(k, x)$ *defined for all* $k \in \mathfrak{S}_N$, *all* $x \in \mathfrak{R}^n$, *and a real-valued function* $\phi(k)$ *defined for all* $k \in \mathfrak{S}_N$, *such that hypotheses* (i) *and* (ii) *in Theorem* 8 *are satisfied, and in addition*

$$\sum_{j \in \mathfrak{S}_{N-1}} \varphi(j) < \min_{\beta \leqq \|x\| \leqq \alpha} V(k_0 + N, x) - V^{\alpha}_{M}(k_0).$$

Furthermore, if f *in* (47) *is "invertible," the converse holds.*

Remark 15. Sufficiency conditions for finite-time stability of (47) were first obtained by Michel and Wu [35].

Remark 16. Necessary and sufficient conditions for finite-time stability of (47) have also been derived by Heinen [36], but in a form different from that presented here.

11. Finite-time stability of linear discrete-time systems. The basic linear theory for finite-time stability of discrete-time systems has been developed in [32], which contains general sufficient conditions, general necessary conditions, as well as results on mean-square finite-time stability under white noise sequence perturbations.

Our objective in this section is to indicate how the Hermite–Fujiwara form of the Schur–Cohn criterion for asymptotic stability of linear constant-coefficient difference equations can be used to obtain a computationally simple test for finite-time stability of linear time-invariant systems. Our exposition of the classical result follows that of Kalman [37].

First we characterize finite-time stability for linear discrete-time systems.

Let F be a real $n \times n$ matrix. $\{\lambda(F)\}$ is the set of eigenvalues of F. If the latter are real, $\tilde{\lambda} = \max \{\lambda(F)\}$. Define the *spectral norm* of F as

$$\|F\|^* = \sqrt{\tilde{\lambda}(F'F)}.$$

Now consider the system of linear equations

$$x(k+1) = A(k)x(k), \quad k = k_0, k_0 + 1, k_0 + 2, \cdots, \tag{50}$$

where $A(k)$ is $n \times n$.

The solution at the lth instant starting with initial state x_0 at time k_0 is

$$x(l; k_0, x_0) = \Omega(l-1, k_0)x_0, \tag{51}$$

where Ω is defined by (3). From (51) and Definition 13 we obtain the following result.

THEOREM 11 (Weiss and Lee [33]).[1] *The system* (51) *is stable* (*Definition* 13) *if and only if*

$$\|\Omega(k, k_0)\|^* < \beta/\alpha, \qquad k = 1, \cdots, N. \tag{52}$$

[1] In [33], $\|x_0\|$ cannot be equal to α in the definition of finite-time stability. Hence (52) and (53) differ from the corresponding conditions in [33] by being strict inequalities.

COROLLARY 4. *If* $A(\cdot)$ *in* (50) *is constant with time, then* (51) *is stable* (*Definition* 13) *if and only if*

$$\|A^k\|^* < \beta/\alpha, \qquad k = 1, \cdots, N. \tag{53}$$

Now, let A in (50) be a constant matrix, and denote the characteristic polynomial of A by

$$P(z) = z^n + \alpha_1 z^{n-1} + \cdots + \alpha_n. \tag{54}$$

Then the constant system (50) is (classically) asymptotically stable if and only if all the zeros of $P(z)$ lie inside the unit circle on the complex plane.

A criterion for $P(z)$ to have all its zeros inside the unit circle was given by Fujiwara [38], using a classical technique of Hermite, as follows.

Let P^* be the polynomial defined by

$$P^*(z) = z^n P(z^{-1}), \tag{55}$$

and define

$$Q(z, w) = \left[\frac{P(z)P^*(w) - P(w)P^*(z)}{z - w}\right] = \sum_{i,j=1}^{n} z^{i-1}\psi_{ij}w^{j-1}. \tag{56}$$

By (55), Q also has the representation

$$Q(z, w) = w^{n-1}\left[\frac{P(z)P(w^{-1}) - P^*(w^{-1})P^*(z)}{zw^{-1} - 1}\right] = w^{n-1}\sum_{i,j=1}^{n} z^{i-1}\varphi_{ij}w^{1-j}. \tag{57}$$

The matrices $\Phi = (\varphi_{ij})$ and $\Psi = (\psi_{ij})$ are $n \times n$ symmetric matrices and are related by $\psi_{ij} = \varphi_{i,n-j}$. In fact, from (54), (55), and (56), one can easily compute φ_{ij} as

$$\varphi_{ij} = \sum_{k=1}^{\min(i,j)} (\alpha_{i-k}\alpha_{j-k} - \alpha_{n-i+k}\alpha_{n-j+k}), \qquad i, j = 1, \cdots, n. \tag{58}$$

THEOREM 12 (Fujiwara). *The zeros of the polynomial* P *in* (54) *lie inside the unit circle on the complex plane if and only if the matrix* Φ *defined by* (58) *is positive definite.*

To apply this to the finite-time stability problem, we need only consider the simple proposition below.

PROPOSITION 3. *If* z_0 *is an eigenvalue of a matrix* F, *then* cz_0 *is an eigenvalue of* cF.

Now, consider the system (50) with A a constant $n \times n$ matrix, and let the characteristic polynomial of $(\alpha^2/\beta^2)(A'^kA^k)$ be given by

$$P_k(z) = z^n + \alpha_{1k}z^{n-1} + \cdots + \alpha_{nk}. \tag{59}$$

Let $\Phi_k = (\varphi_{ij}^{(k)})$ be the $n \times n$ matrix defined by

$$\varphi_{ij}^{(k)} = \sum_{l=1}^{\min(i,j)} (\alpha_{i-l,k}\ \alpha_{j-l,k} - \alpha_{n-i+l,k}\ \alpha_{n-j+l,k}), \qquad i, j = 1, \cdots, n. \tag{60}$$

Then, from Corollary 4, Theorem 12 and Proposition 3, we have the main result of the section.

THEOREM 13. *The system* (50), *with A constant, is stable* (*Definition* 13) *if and only if the matrices* Φ_k *defined by* (60) *are positive definite for* $k = 1, 2, \cdots, N$.

Using a result of Parks [39], an alternative form of Theorem 13 can be obtained. Let P_k be the characteristic polynomial of $(\alpha^2/\beta^2)(A'^kA^k)$ as in (59). Let $P_k^*(x) = z^nP_k(z^{-1})$. Define the n-vector $q_k = \operatorname{col}(q_{k1}, \cdots, q_{kn})$ by the expression

$$\alpha_{kn}P_k(z) - P_k^*(z) = q_{kn}z^{n-1} + \cdots + q_{k1}.$$

Let Ψ_k denote the companion matrix

$$\Psi_k = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \\ & & & \ddots \\ 0 & 0 & \cdots & 1 \\ -\alpha_{kn} & \cdot & \cdots & -\alpha_{k1} \end{bmatrix}.$$

Then we have the following theorem.

THEOREM 14. *The system* (50), *with A constant, is stable* (*Definition* 13) *if and only if there exists a sequence* $\{S_k\}$ *of symmetric, positive definite matrices which satisfy the equations*

$$\Psi_k'S_k\Psi_k - S_k = -q_k'q_k, \qquad k = 1, \cdots, N. \tag{61}$$

REFERENCES

[1] W. FELLER, *An Introduction to Probability Theory and its Applications*, vol. I, John Wiley, New York, 1950.

[2] E. I. JURY, *Theory and Application of the Z-Transform Method*, John Wiley, New York, 1964.

[3] G. BOOLE, *Calculus of Finite Differences*, 4th ed., Chelsea, New York, 1958.

[4] L. M. MILNE-THOMSON, *The Calculus of Finite Differences*, Macmillan, London, 1933.

[5] P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1962.

[6] J. R. RAGAZZINI AND G. FRANKLIN, *Sampled-Data Control Systems*, McGraw-Hill, New York, 1958.

[7] S. GOLDBERG, *Introduction to Difference Equations*, John Wiley, New York, 1958.

[8] E. I. JURY AND YA. Z. TSYPKIN, *Theory of Discrete Automatic Systems* (*Review*), Avtomat. i Telemekh., (1970), pp. 57–81.

[9] L. WEISS, *Controllability for various linear and nonlinear system models*, Seminar on Differential Equations and Dynamical Systems. II, Lecture Notes in Mathematics, Springer-Verlag, New York, 1970, pp. 250–261.

[10] ———, *On the controllability of delay-differential systems*, this Journal, 5 (1967), pp. 575–587.

[11] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, John Wiley, New York, 1955.

[12] R. E. KALMAN, *On the general theory of control systems*, Proc. First IFAC, Butterworth's, London, 1960, pp. 481–492.

[13] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[14] L. WEISS, *Lectures on controllability and observability*, C.I.M.E. Seminar on Controllability and Observability, Edizioni Cremonese, Rome, 1969, pp. 205–289.

[15] E. B. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–38.

[16] H. HERMES, *Controllability and the singular problem*, this Journal, 3 (1965), pp. 241–260.

[17] L. WEISS, *On the structure theory of linear differential systems*, this Journal, 6 (1968), pp. 659–680.

[18] R. E. KALMAN, *Realization theory for nonconstant linear systems*, unpublished notes.

[19] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.
[20] L. WEISS AND R. E. KALMAN, *Contributions to Linear System Theory*, Internat. J. Engrg. Sci., 3 (1964), pp. 141–171.
[21] D. C. YOULA, *The synthesis of linear dynamical systems from prescribed weighting patterns*, SIAM J. Appl. Math., 14 (1966), pp. 527–549.
[22] B. L. HO AND R. E. KALMAN, *Effective construction of state-variable models from input/output functions*, Regelungstechnik, 12 (1966), pp. 545–548.
[23] R. E. KALMAN, *Canonical structure of linear dynamical systems*, Proc. Nat. Acad. Sci., 48 (1962), pp. 596–600.
[24] L. SILVERMAN AND H. MEADOWS, *Equivalent realizations of linear systems*, SIAM J. Appl. Math., 17 (1969), pp. 393–408.
[25] W. BELFIELD AND L. WEISS, *Realization theory with applications to the design of digital filters*, to appear.
[26] F. R. GANTMACHER, *The Theory of Matrices. I, II*, Chelsea, New York, 1959.
[27] R. E. KALMAN, *On partial realizations of a linear input/output map*, Aspects of Network and System Theory, Holt, Rinehart and Winston, New York, 1970.
[28] A. J. TETHER, *Construction of minimal linear state-variable models from finite input/output data*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 427–436.
[29] W. HAHN, *Über die Anwendung der Methode von Ljapunov auf Differenzengleichungen*, Math. Ann., 136 (1958), pp. 430–441.
[30] ———, *Stability of Motion*, Springer-Verlag, New York, 1967.
[31] L. WEISS AND L. LAM, *Stability of nonlinear discrete-time systems*, Internat. J. Control, to appear.
[32] L. WEISS AND J. S. LEE, *Finite-time stability of linear discrete-time systems*, Avtomat. i Telemekh., to appear.
[33] A. A. KAYANDE, *A theorem on contractive stability*, to appear.
[34] J. HURT, *Some stability theorems for difference equations*, SIAM J. Numer. Anal., 4 (1967), pp. 582–596.
[35] A. N. MICHEL AND S. H. WU, *Stability on discrete systems over a finite interval of time*, Internat. J. Control, 9 (1969), pp. 679–693.
[36] J. A. HEINEN, *Quantitative stability of discrete systems*, Michigan Math. J., 17 (1970), pp. 211–216.
[37] R. E. KALMAN, *On the Hermite-Fujiwara theorem in stability theory*, Quart. Appl. Math., 23 (1965), pp. 279–282.
[38] M. FUJIWARA, *Über die Algebraischen Gleichungen, deren Wurzeln in einem Kreise oder in einer Halbene Liegen*, Math. Z., 24 (1926), pp. 160–169.
[39] P. C. PARKS, *Lyapunov and the Schur-Cohn stability criterion*, IEEE Trans. Automatic Control, AC-8 (1964), p. 121.
[40] H. W. SORENSON, *Controllability and observability of linear, stochastic, time-discrete control systems*, Advances in Control Systems, C. T. Leondes, ed., vol. 6, Academic Press, New York, 1968.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

INVARIANT DESCRIPTION OF LINEAR, TIME-INVARIANT CONTROLLABLE SYSTEMS*

V. M. POPOV†

Abstract. One considers a class of linear, time-invariant controllable systems, together with a group of transformations of the systems, and one determines a complete system of independent invariants. One examines also how these invariants are influenced by a feedback. The results are also interpreted from the point of view of the canonical forms and one shows that—under some natural hypotheses—the canonical forms contained (implicitly) in the paper are described in terms of a minimal number of parameters. The algebraic concept of "universal element" is a principal tool in proving the last result and a leading guide for the whole paper.

1. Introduction. In this section we discuss the general ideas which guide the present paper. Sections 2 and 3 of the paper can be read independently of this Introduction. However, in §4 we return to the interpretation of the present section.

Consider the linear, time-invariant control system described by a differential equation of the form (see Remark 1.1)

$$\dot{x} = Ax + Bu,$$

where $x(t) \in F^n$, $A \in F^{n \times n}$, $B \in F^{n \times m}$ and F is a field (one may consider only the field of real numbers). We assume that the properties we study remain unchanged if one performs a transformation of the form $\tilde{x} = Rx$, where $R \in F^{n \times n}$ is a nonsingular matrix. Then (1) becomes $\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}u$, where $\tilde{A} = RAR^{-1}$ and $\tilde{B} = RB$. Instead of considering the equations of the form $\dot{x} = Ax + Bu$ we shall consider the set of all pairs of matrices $(A, B) \in F^{n \times (n+m)}$ (where n and m are fixed, positive integers) together with the set of all transformations of the form $\tilde{A} = RAR^{-1}$, $\tilde{B} = RB$ (for all nonsingular matrices $R \in F^{n \times n}$).

Since R is arbitrary, an obvious problem is to choose this matrix so as to simplify as much as possible the study of the considered system. In the existing literature, this problem is solved (at least partially) by introducing some "canonical" forms (see Remark 1.2). The basic idea of this approach may be visualized by the commutative diagram in Fig. 1, where $F^{n \times (n+m)}$ is the set of all pairs of matrices (A, B) with the indicated dimensions, C is a set of so-called "canonical forms" and y is a function which allows one to determine, for every pair (A, B), a corresponding canonical form in C. Observe that any property of the pair (A, B) can be described by a function f: $F^{n \times (n+m)} \to S$, where S is a suitable set. As visualized by the diagram in Fig. 1, after introducing the canonical forms, the study of the original function f is replaced by the study of a new function h (which is simpler).

According to this interpretation, the introduction of the canonical forms appears as an attempt to solve a problem of universality (see, for instance, S.

* Received by the editors September 21, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Department of Electrical Engineering, University of Maryland, College Park, Maryland 20742. This research was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-69-1646.

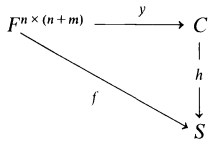


FIG. 1

MacLane and G. Birkhoff [1] or N. Bourbaki [2]). To formulate precisely the problem, let us first introduce a class of functions: For every set S, let $W(S)$ be the set of all the functions $f\colon F^{n\times(n+m)}\to S$, $((A,B)\mapsto f(A,B))$, such that, for every nonsingular matrix $R\in F^{n\times n}$, one has $f(RAR^{-1},RB)=f(A,B)$. Then our problem of universality can be stated as follows: Find a set C and a function $y\in W(C)$ such that the following property is true: For every set S and every function $f\in W(S)$ there exists exactly one function $h\colon C\to S$ such that $f=h\circ y$ (that is, the diagram in Fig. 1 is commutative).

Using the terminology from [1], a pair (y, C) with the above property is called a universal element (for the functor—from the category of sets to the category of sets—having as object function the function $S\mapsto W(S)$, introduced above; this is a subfactor of the exponential functor $(\cdot)^X$, where $X=F^{n\times(n+m)}$).

The above problem has a standard solution (see [1]). Introduce the relation E defined as follows: $(\tilde{A},\tilde{B})$ is in the relation E with (A,B) if and only if there exists a nonsingular matrix $R\in F^{n\times n}$ such that $\tilde{A}=RAR^{-1}$ and $\tilde{B}=RB$. Clearly E is an equivalence relation in $F^{n\times(n+m)}$. Take C as the quotient set $F^{n\times(n+m)}/E$ and y as the projection $F^{n\times(n+m)}\to F^{n\times(n+m)}/E$. Then (y, C) is a universal element for our problem. Moreover, this solution is (essentially) unique, in the sense made precise by the uniqueness theorem for universal elements (see [1]).

However, to make the solution amenable to numerical computations, one has to introduce some restrictions upon y and C. One would prefer to obtain C as a subset of F^N (for some integer N which should be as small as possible) and one would like the function y to be as simple as possible (for instance, a rational function). If one introduces such restrictions one obtains in fact a new problem of universality, requiring a new solution. (The problem could even have no solution if one introduces too many conditions for y and C.) Since there exists no general method for solving universal problems, one has to take a direct approach. Observe first that for any solution (y, C) of the universal problem stated above, the following properties are true:

Invariance. That is, $y(RAR^{-1},RB)=y(A,B)$ for every nonsingular matrix $R\in F^{n\times n}$.

Independence. For every $s\in C$, there exists a pair $(A,B)\in F^{n\times(n+m)}$ such that $y(A,B)=s$.

Completeness. If for two pairs $(\tilde{A},\tilde{B})$ and (A,B) from $F^{n\times(n+m)}$ one has $y(\tilde{A},\tilde{B})=y(A,B)$, then there exists a nonsingular matrix $R\in F^{n\times n}$ such that $\tilde{A}=RAR^{-1}$ and $\tilde{B}=RB$.

Indeed, the property of invariance follows from $y\in W(C)$ and from the definition of $W(C)$. To prove the property of independence, suppose there exists

$s_0 \in C$ such that, for every $(A, B) \in F^{n \times (n+m)}$, one has $y(A, B) \neq s_0$; then $h(s_0)$ can be defined arbitrarily and the diagram drawn above remains commutative—contrary to the uniqueness of h. Finally, if two pairs $(\tilde{A}, \tilde{B})$ and (A, B) are not related by any relation of the form $\tilde{A} = RAR^{-1}$, $\tilde{B} = RB$, one can always find a set S and a function $f \in W(S)$ such that $f(\tilde{A}, \tilde{B}) \neq f(A, B)$ and therefore our diagram is not commutative if the property of completeness is not satisfied.

Conversely, if for a pair (y, C) such that $y: F^{n \times (n+m)} \to C$, the properties of invariance, independence and completeness are satisfied, then (y, C) is a universal element for our problem. Indeed, for any set S and any function $f \in W(S)$ one can define a function $h \in W(C)$ as follows: For every $s \in C$, choose (A, B) such that $y(A, B) = s$ (using the property of independence) and define $h(s) = f(A, B)$ (the property of completeness implies that this definition be independent of the particular choice of the pair (A, B)). Then the property of invariance implies that $y \in W(C)$, the relation $f = h \circ y$ follows from the definition of h and the uniqueness of h follows from the property of independence.

The above comments show that, in order to solve our universal problem, it is sufficient to look for a pair (y, C) that satisfies the properties of invariance, independence and completeness.

Guided by this conclusion, we shall solve, in § 2, the following problem: For every pair $(A, B) \in F^{n \times (n+m)}$ find a positive integer N and a vector $y(A, B) \in F^N$ whose coordinates constitute a complete system of independent invariants for the pair (A, B) with respect to the transformations $\tilde{A} = RAR^{-1}$, $\tilde{B} = RB$ (for arbitrary nonsingular matrices $R \in F^{n \times n}$). The corresponding problem of universality will be discussed further in § 4.

Remark 1.1. The results of this paper also remain valid for larger systems, which contain, besides the equation $\dot{x} = Ax + Bu$, other equations involving x and u (for instance, a nonlinear feedback of the form $u = f(x)$).

Remark 1.2. A discussion of many canonical forms can be found in a paper by D. G. Luenberger [3]. The earlier paper of P. Brunovsky [4] introduces one of the most useful forms. However, all these forms are not invariant (except in degenerate cases). Canonical forms which are invariant have been introduced in the (unpublished) paper [5] and some of the results have been mentioned without proof in [6]. A proof of these results can be derived from the proofs of the present paper (which are much shorter and simpler than the proofs from [5]).

2. Determination of a complete system of independent invariants. Let us introduce the following assumptions (see Remark 2.1):

$$\text{rank } (B, AB, \cdots, A^{n-1}B) = n, \tag{1}$$

$$\text{rank } B = m. \tag{2}$$

Consider the ordered set of vectors

$$b_1, b_2, \cdots, b_m, \quad Ab_1, Ab_2, \cdots, Ab_m, \quad A^2b_1, \cdots, A^2b_m, \cdots, \tag{3}$$

where the b_i are the columns of matrix B, written as $B = (b_1 \quad b_2 \quad \cdots \quad b_m)$.

We shall say that a vector $A^k b_j$ from (3) is an *antecedent* of another vector $A^p b_q$ from (3) if and only if $A^k b_j$ is situated before $A^p b_q$ in (3) (that is, if and only if $km + j < pm + q$).

We first introduce some well-known invariants for our problem (see Remark 2.2).

DEFINITION 1. For every integer $i \in \{1, \cdots, m\}$, one defines the ith Kronecker invariant, n_i, as the smallest positive integer such that the vector $A^{n_i}b_i$ is a linear combination of its antecedents.

For any integers $p \geqq 0$ and $q \in \{1, \cdots, m\}$, let Ant (p, q) denote the set of all the antecedents of the vector $A^p b_q$ (that is, the set of all the vectors $A^k b_j$ from (3) such that $km + j < pm + q$). Given an arbitrary set $M \subset F^n$, let us denote by $[M]$ the subspace of F^n spanned by M. Then the conditions from Definition 1 are equivalent to: $A^{n_i - 1}b_i \notin [\text{Ant}\,(n_i - 1, i)]$ and $A^{n_i}b_i \in [\text{Ant}\,(n_i, i)]$. The last condition implies $A^k b_i \in [\text{Ant}\,(k, i)]$ for every $k \geqq n_i$. Therefore the Kronecker invariants from Definition 1 can be characterized by the condition

$$A^k b_i \in [\text{Ant}\,(k, i)] \quad \text{if and only if} \quad k \geqq n_i. \tag{4}$$

DEFINITION 2. A vector $A^k b_j$ from (3) is called *regular* if and only if $k < n_j$. The set of all the regular vectors from (3) will be denoted by Reg.

Now condition (4) can be improved as follows.

PROPOSITION 1. *For every $i = 1, 2, \cdots, m$ and every integer $k \geqq n_i$, one has*

$$A^k b_i \in [\text{Ant}\,(k, i) \cap \text{Reg}]. \tag{5}$$

In words: every nonregular vector from (3) *is a linear combination of its regular antecedents.*

Proof. Condition (5) is true for the first nonregular vector from (3). Indeed, this vector is necessarily of the form $A^{n_r}b_r$ for some integer r (more precisely, for the unique integer r that satisfies the inequalities $n_r m + r \leqq n_s m + s$, for every $s = 1, 2, \cdots, m$). From (4) one obtains $A^{n_r}b_r \in [\text{Ant}\,(n_r, r)]$ and since $A^{n_r}b_r$ is the first nonregular vector from (3), all its antecedents are regular, i.e., Ant $(n_r, r) \subset$ Reg. The relations obtained imply (5) for $k = n_r$ and $i = r$.

Suppose now, by induction, that (5) is true for every nonregular antecedent of $A^p b_q$ (that is, for every vector $A^k b_i$ such that $km + i < pm + q$ and $k \geqq n_i$) and assume also that $p \geqq n_q$. Then $[\text{Ant}\,(k, i)] \subset [\text{Ant}\,(p, q)]$ and therefore one obtains from (5), under the hypothesis of induction,

$$A^k b_i \in [\text{Ant}\,(p, q) \cap \text{Reg}] \tag{6}$$

for every vector $A^k b_i \in \text{Ant}\,(p, q)$ such that $A^k b_i \notin \text{Reg}$. On the other hand, if $A^k b_i \in \text{Ant}\,(p, q)$ and $A^k b_i \in \text{Reg}$, condition (6) is obviously true. Thus (6) is true for every $A^k b_i \in \text{Ant}\,(p, q)$. This implies

$$[\text{Ant}\,(p, q)] \subset [\text{Ant}\,(p, q) \cap \text{Reg}]. \tag{7}$$

Since $p \geqq n_q$ one obtains, from (4), that $A^p b_q \in [\text{Ant}\,(p, q)]$. This and (7) prove (5) for $k = p$ and $i = q$. Thus (5) is proved by induction.

PROPOSITION 2. *The regular vectors are linearly independent.*

Indeed, consider the equation

$$\sum_{j=1}^{m} \sum_{k=1}^{n_j - 1} \rho_{jk} A^k b_j = 0. \tag{8}$$

If the coefficients ρ_{jk} do not all vanish, let $A^p b_q$ be the *last vector from* (3) with the properties $p \leqq n_q - 1$ and $\rho_{qp} \neq 0$. Then (8) implies that $A^p b_q \in [\mathrm{Ant}\,(p, q)]$ and hence (see (4)), $p \geqq n_q$, a contradiction.

PROPOSITION 3. *The Kronecker invariants from Definition* 1 *satisfy the relation*

$$n_1 + n_2 + \cdots + n_m = n. \tag{9}$$

Proof. From (5) it follows that every vector from (3) is a linear combination of regular vectors. Using (1), one sees that the set (3) contains exactly n linearly independent vectors. From Proposition 2 it follows that the number of the regular vectors equals n. Using Definition 2, one obtains (9).

Now we can define other quantities which (as proved later) are invariant.

COROLLARY 1. *There exists exactly one set of ordered numbers* $\alpha_{ijk} \in F$, *defined* for $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, i - 1$, $k = 0, 1, \cdots, \min(n_i, n_j - 1)$ *and for* $i = 1, 2, \cdots, m$, $j = i, i + 1, \cdots, m$, $k = 0, 1, \cdots, \min(n_i, n_j) - 1$ *such that, for every* $i = 1, 2, \cdots, m$, *one has*

$$A^{n_i} b_i = \sum_{j=1}^{i-1} \sum_{k=0}^{\min(n_i, n_j - 1)} \alpha_{ijk} A^k b_j + \sum_{j=i}^{m} \sum_{k=0}^{\min(n_i, n_j) - 1} \alpha_{ij} A^k b_j, \tag{10}$$

where n_i *are the Kronecker invariants from Definition* 1.

Proof. The existence and uniqueness of the numbers α_{ijk} from (10) follows immediately from (5), using also Proposition 2.

Now we show, step by step, that the numbers n_i and α_{ijk}, defined in Definition 1 and in Corollary 1, constitute a complete system of independent invariants for our problem.

PROPOSITION 4. *The numbers* n_i *and* α_{ijk} (*Definition* 1, *Corollary* 1) *remain unchanged if* (A, B) *is replaced by* $(\tilde{A}, \tilde{B}) = (RAR^{-1}, RB)$ $(\det R \neq 0)$.

Proof. Indeed, n_i and α_{ijk} have been defined above using only the vectors $A^k b_j$ from the set (3). If the pair (A, B) is replaced by $(\tilde{A}, \tilde{B}) = (RAR^{-1}, RB)$, then the vectors $A^k b_j$ are replaced by $\tilde{A}^k \tilde{b}_j = (RAR^{-1})^k R b_j = RA^k b_j$. Thus all we have to do is to multiply, on the left, every vector $A^k b_j$, from all the preceding proofs, with the nonsingular matrix R. It is easy to see that the numbers n_i and α_{ijk} remain unchanged.

PROPOSITION 5. *The set of the invariants* n_i, α_{ijk} *is independent. In other words, given arbitrarily the positive integers* $n_i, i = 1, 2, \cdots, m$ (*satisfying* (9)), *and the numbers* $\alpha_{ijk} \in F$ (*where the indices* i, j *and* k *take the values specified in Corollary* 1), *there exists a pair of matrices* (A, B) *whose invariants* (*defined as in Definition* 1 *and Corollary* 1) *are precisely the above numbers* n_i *and* α_{ijk}.

Proof. Choose arbitrarily a basis of F^n and denote the vectors of this basis as $e_{10}, e_{11}, e_{12}, \cdots, e_{1,n_1-1}, e_{20}, e_{21}, \cdots, e_{2,n_2-1}, \cdots, e_{m0}, e_{m1}, \cdots, e_{m,n_m-1}$ (this is possible because of (9)). Denote by N the corresponding matrix

$$N = (e_{10} \quad e_{11} \quad \cdots \quad e_{1,n_1-1} \quad \cdots \quad e_{m0} \quad e_{m1} \quad \cdots \quad e_{m,n_m-1}) \tag{11}$$

which is obviously nonsingular. Define the columns b_i of the matrix B as

$$b_i = e_{i0}, \qquad i = 1, 2, \cdots, m. \tag{12}$$

To define A it is sufficient to define the matrix AN (since N is nonsingular) or its columns Ae_{ij}. Define these vectors as

$$Ae_{ij} = e_{i,j+1} \quad \text{for } i = 1, 2, \cdots, m, \quad j = 0, 1, \cdots, n_i - 2, \tag{13}$$

$$Ae_{i,n_i-1} = \sum_{j=1}^{i-1} \sum_{k=0}^{\min(n_i,n_j)-1} \alpha_{ijk} e_{jk} + \sum_{j=1}^{m} \sum_{k=0}^{\min(n_i,n_j)-1} \alpha_{ijk} e_{jk}. \tag{14}$$

It is easy to check that, for the pair (A, B) defined above, conditions (1) and (2) are satisfied. It remains to show that the invariants of the pair (A, B) are precisely the numbers n_i and α_{ijk}. From (13) and (12) one obtains $e_{ij} = Ae_{i,j-1} = A^2 e_{i,j-2} = \cdots$, and hence,

$$e_{ij} = A^j b_i, \qquad i = 1, 2, \cdots, m, \qquad j = 0, 1, \cdots, n_i - 1. \tag{15}$$

From (14) and (15) one obtains (10), which implies

$$A^{n_i} b_i \in [\text{Ant}\,(n_i, i)] \quad \text{for } i = 1, 2, \cdots, m. \tag{16}$$

Let $\tilde{n}_i$ be the Kronecker invariants, obtained as in Definition 1 for our pair (A, B). Then from (16) it follows that $n_i \geqq \tilde{n}_i$. Hence, using (9) and the similar relation $\tilde{n}_1 + \tilde{n}_2 + \cdots + \tilde{n}_m = n$ (see Proposition 3) one obtains $\tilde{n}_i = n_i$ for $i = 1, 2, \cdots, m$. Now (10) (obtained before in the proof) shows that the α_{ijk} constitute the invariants, given by Corollary 1, for our pair (A, B). See also Remark 2.3.

PROPOSITION 6. *The set of the invariants* n_i (*Definition* 1), α_{ijk} (*Corollary* 1) *is complete. In other words, if for two pairs of matrices* (A, B) *and* $(\tilde{A}, \tilde{B})$ (*of corresponding dimensions*) *the invariants, given by Definition* 1 *and Corollary* 1, *coincide, there exists a nonsingular matrix* $R \in F^{n \times n}$ *such that* $\tilde{A} = RAR^{-1}$ *and* $\tilde{B} = RB$.

Proof. From Propositions 2 and 3, using also Definition 2, it follows that the matrices

$$U = (b_1 \quad Ab_1 \cdots A^{n_1-1} b_1 \quad b_2 \cdots A^{n_2-1} b_2 \cdots b_m \cdots A^{n_m-1} b_m), \tag{17}$$

$$\tilde{U} = (\tilde{b}_1 \quad \tilde{A}\tilde{b}_1 \cdots \tilde{A}^{n_1-1} \tilde{b}_1 \quad \tilde{b}_2 \cdots \tilde{A}^{n_2-1} \tilde{b}_2 \cdots \tilde{b}_m \cdots \tilde{A}^{n_m-1} \tilde{b}_m) \tag{18}$$

are nonsingular (we have also used the condition that the invariants n_i are identical for the two pairs of matrices (A, B) and $(\tilde{A}, \tilde{B})$). Therefore, one can define $R = \tilde{U} U^{-1}$, which implies

$$\tilde{U} = RU, \tag{19}$$

$$\tilde{A}^k \tilde{b}_j = RA^k b_j, \quad j = 1, 2, \cdots, m, \quad k = 0, 1, \cdots, n_j - 1 \tag{20}$$

(see (17) and (18)). From (10) and from the similar equation, written for the pair $(\tilde{A}, \tilde{B})$, one obtains that

$$\tilde{A}^{n_j} \tilde{b}_j = RA^{n_j} b_j, \qquad j = 1, 2, \cdots, m \tag{21}$$

(here we have used the condition that the invariants α_{ijk} are identical for the two pairs of matrices). From (20), for $k = 0$, one obtains $\tilde{B} = RB$. From (20) and (21) it follows that $\tilde{A}\tilde{B} = RAU$ (see (17) and (18)). Hence $\tilde{A} = RAU\tilde{U}^{-1} = RAR^{-1}$ (see (19)).

We have proved the following theorem, which summarizes the results of this section.

THEOREM 1. *If conditions* (1) *and* (2) *are satisfied, the numbers* n_i (*Definition* 1) *and* α_{ijk} (*Corollary* 1) *constitute a complete system of independent invariants for the pair* (A, B) *with respect to the transformations of the form* $\tilde{A} = RAR^{-1}$, $\tilde{B} = RB$, *for arbitrary nonsingular matrices* $R \in F^{n \times n}$.

Remark 2.1. Condition (1) means that the pair (A, B) is completely controllable (see R. E. Kalman [7]). Condition (2) obviously implies $m \leqq n$. Conditions (1) and (2) are introduced in order to simplify the study and obtain what could be called a "pure control problem." (In the particular case in which B is absent our problem becomes equivalent to a well-known classical problem, solved by the Jordan canonical forms.)

Condition (2) can be easily eliminated at the price of some notational complications. Then some of the Kronecker invariants from Definition 1 may also take the value 0. Condition (1) was used only in order to show that our system of invariants is complete (Propositions 3 and 6). Therefore if the pair (A, B) is not completely controllable, the numbers n_i and α_{ijk} from Definition 1 and Corollary 1 still represent a system of independent invariants for our problem, but other invariants must be found in order to obtain a *complete* system of such invariants. The complete theory obviously contains, in particular, the classical theory of Jordan forms—and requires more space to be developed.

Remark 2.2. The numbers n_i from Definition 1 are implicitly contained in Kronecker's theory of singular pencils of matrices (see, for instance, F. R. Gantmacher [18])—as recent results of H. H. Rosenbrock [9] and R. E. Kalman [10] have pointed out. The same numbers occur in the classification of controllable systems, established by P. Brunovsky [11], even in the time-varying case. We note, however, that the order of the numbers n_i in Definition 1 differs from the usual one. We also remark that many stages of our proofs can be found, in slightly different forms, in the existing literature (for instance, in [3], [4], [10]).

Remark 2.3. The proof of the property of independence (Proposition 5) also solves the problem of determining a canonical form—that is, a pair of matrices (A, B) which is determined only in terms of the invariants n_i, α_{ijk} and has the above invariants. (First one calculates the invariants, as in Definition 1 and Corollary 1, for the given pair of matrices, and then one determines the corresponding canonical form as in the proof of Proposition 5.) There are other canonical forms worthy of notice which can be obtained similarly.

3. Other properties of the invariants. The following form of the results of the previous section (closer to the results from the unpublished paper [5]) is useful for some applications (see, for instance, Theorem 3 below).

THEOREM 2. *Let* (A, B) *be a pair of matrices for which conditions* (1) *and* (2) *are satisfied. Let* n_i, $i = 1, 2, \cdots, m$, *be some positive integers satisfying the condition* $n_1 + n_2 + \cdots + n_m = n$. *Let* $\alpha_{ijk} \in F$ *be some numbers, defined for* $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, i - 1$, $k = 0, 1, \cdots, \min(n_i, n_j - 1)$ *and for* $i = 1, 2, \cdots, m$, $j = i, \cdots, m$, $k = 0, 1, \cdots, \min(n_i, n_j) - 1$. *Then the following properties of the above numbers* n_i *and* α_{ijk} *are equivalent*:

(i) *The numbers* n_i *and* α_{ijk} *represent the complete system of independent invariants determined as in Definition* 1 *and Corollary* 1 (*see also Theorem* 1).

(ii) *There exists a set of vectors* $s_{ij} \in F^n$, $i = 1, 2, \cdots, m, j = 0, 1, \cdots, n_i - 1$,

such that the matrix

$$M = (s_{10} \quad s_{11} \cdots s_{1,n_1-1} \quad s_{20} \cdots s_{2,n_2-1} \cdots s_{m0} \cdots s_{m,n_m-1}) \tag{22}$$

is nonsingular and the following equations are satisfied:

$$-As_{i0} = \sum_{j=1}^{m} \alpha_{ij0} b_j, \tag{23}$$

$$s_{i,k-1} - As_{ik} = -\sum_{\substack{j=1 \\ n_j > k}}^{m} \alpha_{ijk} b_j, \qquad k = 1, 2, \cdots, n_i - 1, \tag{24}$$

$$s_{i,n_i-1} = -\sum_{\substack{j=1 \\ n_j > n_i}}^{i-1} \alpha_{ijn_i} b_j + b_i \tag{25}$$

for $i = 1, 2, \cdots, m$. *The symbol* $\sum_{j=1, n_j > k}^{m}$ *from* (24) *means that the summation involves only those integers* $j \in \{1, \cdots, m\}$ *for which* $n_j > k$; *the corresponding symbol from* (25) *has a similar meaning.*

(iii) *There exists a polynomial matrix*

$$S(\sigma) = (s_1(\sigma) \quad s_2(\sigma) \cdots s_m(\sigma)) \tag{26}$$

whose columns $s_i(\sigma)$ *have the form*

$$s_i(\sigma) = s_{i0} + s_{i1}\sigma + \cdots + s_{i,n_i-1}\sigma^{n_i-1}, \tag{27}$$

where the coefficients $s_{ik} \in F^n$ *have the property*

$$\det (s_{10} \quad s_{11} \cdots s_{1,n_1-1} \quad s_{20} \cdots s_{2,n_2-1} \cdots s_{m0} \cdots s_{m,n_m-1}) \neq 0 \tag{28}$$

such that

$$(\sigma I - A)S(\sigma) = BT(\sigma), \tag{29}$$

where I *is the identity matrix in* $F^{n \times n}$ *and* $T(\sigma)$ *is a polynomial matrix whose entries* $(T(\sigma))_{ji}$ *(on the row* j *and column* i*) have the expressions*

$$(T(\sigma))_{ii} = -\sum_{k=0}^{n_i-1} \alpha_{iik}\sigma^k + \sigma^k + \sigma^{n_i}, \qquad i = 1, 2, \cdots, m, \tag{30}$$

$$(T(\sigma))_{ji} = -\sum_{k=0}^{\min(n_i, n_j - 1)} \alpha_{ijk}\sigma^k, \qquad i = 2, 3, \cdots, m, \quad j = 1, 2, \cdots, i-1, \tag{31}$$

$$(T(\sigma))_{ji} = -\sum_{k=0}^{\min(n_i, n_j) - 1} \alpha_{ijk}\sigma^k, \qquad i = 1, 2, \cdots, m-1, \quad j = i+1, \cdots, m. \tag{32}$$

Proof. (i) $\Rightarrow$ (ii). Define s_{i,n_i-1}, for $i = 1, 2, \cdots, m$, by (25) and then define successively $s_{i,n_i-2}, s_{i,n_i-3}, \cdots, s_{i0}$ by (24), for $k = n_i - 1, n_i - 2, \cdots, 1$. Then (23) is also satisfied. Indeed, if for a given i one multiplies every equation in (24), on the left, with A^k and one multiplies also (25), on the left, with A^{n_i} and one adds all the obtained equations, one can rearrange the result as

$$As_{i0} = -\sum_{j=1}^{m} \sum_{k=1}^{\min(n_i, n_j) - 1} \alpha_{ijk} A^k b_j - \sum_{\substack{j=1 \\ n_j > n_i}}^{i-1} \alpha_{ijn_i} A^{n_i} b_j + A^{n_i} b_i. \tag{33}$$

Hence one obtains (23) as a consequence of (10). It remains to show that the matrix M, (22), is nonsingular.

Consider a vector v satisfying the condition $v^T M = 0$ (that is, $v^T s_{ij} = 0$ for $i = 1, 2, \cdots, m$ and $j = 0, 1, \cdots, n_i - 1$; the superscript T denotes transpose). Then from (25) one obtains successively $v^T b_i = 0$ for $i = 1, 2, \cdots, m$, that is, $v^T B = 0$. Therefore from (23) and (24) one obtains $v^T A s_{ik} = 0$ or (see (22)) $v^T A M = 0$. Thus the preceding argument can be applied after replacing v^T by $v^T A$, which gives $v^T AB = 0$ and $v^T A^2 M = 0$. Further application of the same argument gives $v^T A^k B = 0$ for $k = 0, 1, \cdots, n - 1$, which, according to (1), implies $v = 0$. Therefore matrix M is nonsingular.

(ii) $\Rightarrow$ (i). As before, from (24)–(25) one obtains (33), which, taking into account (23), implies (10). The fact that n_i and α_{ijk} are precisely the invariants given by Definition 1 and Corollary 1 is proved as in the last part of the proof of Proposition 5.

(ii) $\Rightarrow$ (iii). This is checked by direct calculations (choosing s_{ij} as in (22)–(25)).

(iii) $\Rightarrow$ (ii) is also verified by direct calculations.

We remark that the principal result from Theorem 2 is the fact that the invariants from § 2 are immediately obtained whenever one has a formula like (29), where S and T have the specified form. This form of the results is particularly helpful in many problems. Consider, as an example, the problem of the "feedback transformations" of the form $\tilde{A} = A - BQ$, $\tilde{B} = B$, where $Q \in F^{m \times n}$. A well-known result states that the coefficients of the characteristic equation of the matrix $A - BQ$ can be arbitrarily modified by a convenient choice of the matrix Q. However, one can do more than that—and the most one can do is expressed, in terms of our invariants, by the following theorem.

THEOREM 3. *Let $(A, B) \in F^{n \times (n+m)}$ satisfy conditions* (1) *and* (2). *Let n_i and α_{ijk} be the corresponding invariants, as in Definition* 1 *and Corollary* 1. *Let $\beta_{ijk} \in F$ be arbitrary numbers, defined for $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, m$, $k = 0, 1, \cdots, \min(n_i, n_j) - 1$. Then there exists a matrix $Q \in F^{m \times n}$ such that the invariants $\tilde{n}_i, \tilde{\alpha}_{ijk}$ of the pair of matrices $(A - BQ, B)$ (Definition* 1, *Corollary* 1) *are given by the relations*

$$\tilde{n}_i = n_i, \quad i = 1, 2, \cdots, m, \tag{34}$$

$$\tilde{\alpha}_{ijk} = \beta_{ijk} \quad \text{for } i, j = 1, 2, \cdots, m, \quad k = 0, 1, \cdots, \min(n_i, n_j) - 1, \tag{35}$$

$$\tilde{\alpha}_{ijk} = \alpha_{ijn_i} \quad \text{for } k = n_i, \quad i = 1, 2, \cdots, m, \tag{36}$$

and for every integer $j \in \{1, \cdots, i - 1\}$ for which $n_j > n_i$.

Proof. According to Theorem 2 one has (29), where S and T have the specified forms. From (29) it immediately follows that $(\sigma I - A - BQ)S(\sigma) = B\tilde{T}(\sigma)$, where $\tilde{T}(\sigma) = T(\sigma) - QS(\sigma)$. Again using Theorem 2 one sees that, if $\tilde{T}(\sigma)$ has the form specified by (30)–(32) (replacing T by $\tilde{T}$ and α_{ijk} by $\tilde{\alpha}_{ijk}$), then n_i and $\tilde{\alpha}_{ijk}$ represent the invariants of the pair $(A - BQ, B)$. Using (27) and (28) one readily sees that, by choosing Q conveniently, one can satisfy (30)–(32), for T replaced by $\tilde{T}$ and α_{ijk} replaced by the numbers $\tilde{\alpha}_{ijk}$ from (35) and (36). Indeed, to accomplish this one must have

$$Qs_{ij} = t_{ij}, \qquad i = 1, 2, \cdots, m, \quad j = 0, 1, \cdots, n_i - 1, \tag{37}$$

where the vectors $t_{ij} \in F^n$ have the form

$$t_{ik} = \begin{pmatrix} \beta_{i1k} - \alpha_{i1k} \\ \beta_{i2k} - \alpha_{i2k} \\ \cdots\cdots\cdots \\ \beta_{imk} - \alpha_{imk} \end{pmatrix}, \qquad i = 1, 2, \cdots, m, \quad k = 0, 1, \cdots, n_i - 1,$$

where we took $\alpha_{ijk} = \beta_{ijk} = 0$ for $k \geqq n_j$. The existence of the matrix Q satisfying (37) follows immediately from (28).

Theorem 3 states, in other words, that, by choosing conveniently, one can modify as one pleases all the invariants of our system, except the invariants which occur in (34) and (36). A little thinking shows that the invariants from (34) and (36) constitute a complete system of independent invariants for the pair (A, B), with respect to the transformations $\tilde{A} = R(A + BQ)R^{-1}$, $\tilde{B} = RB$. One can also see that a complete system of independent invariants with respect to the transformations $\tilde{A} = R(A + BQ)R^{-1}$, $\tilde{B} = RBP^{-1}$ (where $P \in F^{m \times m}$ is an arbitrary non-singular matrix) is given by the *unordered* set of Kronecker invariants n_i. This limit case gives the Brunovsky classes from [11].

4. Canonical forms and invariant descriptions using the minimal number of parameters. In § 2 we found, in fact, a new universal element (y, C) for the problem formulated in § 1. y is the function $(A, B) \mapsto (n_i \alpha_{ijk})$ defined in Definition 1 and Corollary 1. Let K be the set of the matrices $(A, B) \in F^{n \times (n+m)}$ satisfying (1) and (2). The set C is the image of K under the function y and is obviously a subset of F^N, for some integer N. (One can see that the number of invariants given by Definition 1 and Corollary 1 is $\leqq m(n + 1)$; therefore, one may take $N = m(n + 1)$.) The universality of (y, C) follows from the properties of invariance, independence and completeness, as shown in the Introduction.

Now we make more precise the comments from the Introduction. In the proof of Proposition 5, one defines a function r, $(n_i, \alpha_{ijk}) \mapsto r(n_i, \alpha_{ijk}) = (A, B)$ such that $y(A, B) = (n_i, \alpha_{ijk})$. Let L be the image of C under r. Obviously $L \subset K$. Denote by $\hat{f}$ the restriction of the function f to L. Then the diagram in Fig. 2 is

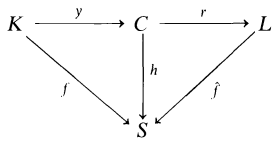


FIG. 2

commutative. Indeed, using Definition 1, Corollary 1 and Proposition 5, one defines successively $(A, B) \mapsto y(A, B) = (n_i, \alpha_{ijk}) \to r(n_i, \alpha_{ijk}) = (\tilde{A}, \tilde{B}) \mapsto \hat{f}(\tilde{A}, \tilde{B})$ (we have denoted by $(\tilde{A}, \tilde{B})$ the pair of matrices given by Proposition 5). One also has $y(\tilde{A}, \tilde{B}) = (n_i, \alpha_{ijk}) = y(A, B)$, which implies (by Proposition 6) $\tilde{A} = RAR^{-1}$ and

$\tilde{B} = RB$. Hence $\hat{f}(\tilde{A}, \tilde{B}) = \hat{f}(RAR^{-1}, RB) = f(A, B)$, since $\hat{f}$ is the restriction of f and $f \in W(S)$ (see the Introduction). On the other hand, since (y, C) is universal, one has $h = \hat{f} \circ r$ (where h is the unique function for which $f = h \circ y$).

From Proposition 5 it follows that r is a bijection. Thus the pair (z, L)—where $z = r \circ y$—is a new universal element for our problem.

Since L is a subset of $F^{n \times (n+m)}$, the elements of L qualify to be called "canonical forms." Moreover, since r is a bijection, one can say that any canonical form from L is "described" by the corresponding invariants $(n_i, \alpha_{ijk}) \in C$. From this point of view, our invariants can be called "parameters." The next question to answer is whether one can find some canonical forms which can be described by a smaller number of parameters.

To make this problem simpler and more precise we shall consider in the following that F represents the field of real numbers. We shall also assume that the (ordered) Kronecker invariants n_i (Definition 1) are fixed. In other words, we consider only those pairs of matrices $(A, B) \in K$, for which the invariants n_i (Definition 1) are equal to some given numbers. Let us denote by $K(n_i)$ this set of pairs of matrices. Obviously $K(n_i) \subset K$.

Let us use the letters y, C, r, L and W in a different sense as before (keeping, however, essentially similar meanings). By y we denote the function $(A, B) \mapsto (\alpha_{ijk})$ described in Corollary 1. C will denote the image of $K(n_i)$ under y. Observe that, if the n_i are given, the number of the invariants α_{ijk} from Corollary 1 is well-determined; let us denote it by $N(n_i)$. Then we have obviously $C = F^{N(n_i)}$ (see Proposition 5). By r we denote the function $(\alpha_{ijk}) \mapsto (A, B)$, given in the proof of Proposition 5. L will represent the image of C under r. From Proposition 5 it follows that $L \subset K(n_i)$. For every set S let us introduce a class $\tilde{W}(S)$ of functions, defined as follows: $\tilde{W}(S)$ is the set of all the functions $f: K(n_i) \to S, ((A, B) \mapsto f(A, B))$, with the property $f(RAR^{-1}, RB) = f(A, B)$ for every nonsingular matrix R. From Proposition 4 it follows that $y \in \tilde{W}(C)$ and since r is a bijection (see Proposition 5) one also has $r \circ y \in \tilde{W}(L)$. With the above interpretation of the symbols, the diagram in Fig. 2 becomes the diagram in Fig. 3, where $\tilde{f}$ is the restriction of f to the set L.

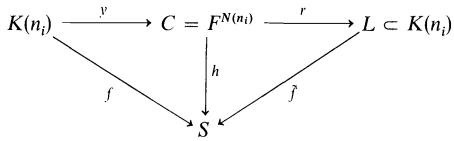


FIG. 3

The function $S \mapsto \tilde{W}(S)$, defined above, is the object function of a functor (a subfunctor of the exponential functor $(\cdot)^{K(n_i)}$). The pair (y, C) is now a universal element for this functor and the same is true for the element $(r \circ y, L)$. The canonical forms are the elements of L and every canonical form is described, using the bijection r, in terms of the parameters α_{ijk}, representing the corresponding element of C.

Observe now that one can say more about the functions y and r. From Corollary 1 and equation (10) it follows that the numbers α_{ijk} (which exist and are finite for every $(A, B) \in K(n_i)$) can be obtained, in the neighborhood from $K(n_i)$ of any particular point (A_0, B_0) from $K(n_i)$, as the quotient of some determinants, involving only the components of the vectors $A^k b_j$. This implies that the function $y: K(n_i) \to C$ is continuous at every point from $K(n_i)$ (considering $K(n_i)$, in the natural way, as a metric subspace of $F^{n \times (n+m)}$ with the standard topology). From Proposition 5 it follows that the function r is also continuous.

Now we ask whether it is possible to find some canonical forms—with all the properties shown above—but described by a smaller number of parameters. In other words, we want to find a similar commutative diagram (shown in Fig. 4),

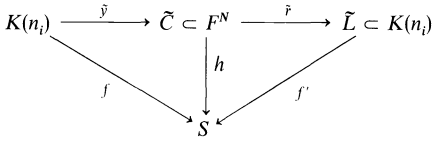


FIG. 4

where $N < N(n_i)$. One assumes, as before, that $\tilde{y} \in \tilde{W}(\tilde{C})$, $f \in \tilde{W}(S)$, f' is the restriction of f to $\tilde{L}$, $\tilde{r}$ is continuous, $(\tilde{y}, \tilde{C})$ is a universal element for the functor defined by $\tilde{W}$ and $\tilde{y}$ is continuous at every point in $K(n_i)$. We now show that, under these assumptions, the condition $N < N(n_i)$ is contradictory. Indeed, using the uniqueness theorem for universal elements, one finds that there exists a bijection $g: C \cong \tilde{C}$ such that $\tilde{y} = g \circ y$ and $y = g^{-1} \circ \tilde{y}$. Moreover, from our last diagram, taking $f = y$ and $S = C$ (which, according to the property of universality for $(\tilde{y}, \tilde{C})$, implies $\tilde{h} = g^{-1}$) one finds that $g^{-1} = f' \circ \tilde{r} = y' \circ \tilde{r}$. Here y' is the restriction of y to $\tilde{L}$ and therefore is continuous. Since $\tilde{r}$ was also assumed to be continuous, the preceding condition $g^{-1} = y' \circ \tilde{r}$ implies that g^{-1} is continuous. Similar arguments show that g is also continuous. Therefore $g: C \cong \tilde{C}$ is a bicontinuous bijection. The classical theorem on dimensional invariance (see L. E. J. Brower [12] or S. Lefschetz [13]) shows that this situation can occur—for $C = F^{N(n_i)}$, $\tilde{C} \subset F^N$ and F, the field of real numbers—only if $N \geqq N(n_i)$.

5. Final comments. The simplicity of the proofs from § 2 indicates that the results can be extended to more general cases. In fact, using basically the same ideas, the time-varying case can be treated quite similarly. Theorem 3 is also only one example of the various possible applications of the invariants in control theory. The universal elements can certainly find a broader use in control theory, as a guide for further progress in conceptual as well as computational problems.

REFERENCES

[1] S. MACLANE AND G. BIRKHOFF, *Algebra*, Macmillan, New York, 1967.

[2] N. BOURBAKI, *Theory of Sets*, Hermann, Paris, 1968.

[3] D. G. LUENBERGER, *Canonical forms for linear multivariable systems*, IEEE Trans. Automatic Control, Short Papers, AC-12 (1967), pp. 290–293.

[4] P. BRUNOVSKY, *On stabilization of linear systems under a certain class of persistent perturbations*, Differential Equations, 2 (1966), pp. 401–405.

[5] V. M. POPOV, *The invariants of the multivariable control systems*, Rep. Power Institute of the Academy of the Socialist Republic of Romania, Bucharest, 1968.
[6] ———, *Some properties of the control systems with irreducible matrix-transfer functions*, Lecture Notes in Mathematics, No. 144, Seminar on Differential Equations and Dynamical Systems. II, Springer-Verlag, Berlin, 1969.
[7] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.
[8] F. R. GANTMACHER, *The Theory of Matrices*, vol. II, Chelsea, New York, 1960.
[9] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, John Wiley-Interscience, New York, 1970.
[10] R. E. KALMAN, *Kronecker invariants and feedback*, Proc. Conference on Ordinary Differential Equations (Mathematics Research Center, Madison, Wis., 1971), Naval Research Laboratory, Washington D.C., 1971.
[11] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika Cislo, 3 (1970), no. 6, pp. 173–188.
[12] L. E. J. BROWER, *Beweis der Invarianz der Dimensionenzahl*, Math. Ann., 70 (1911), pp. 161–165.
[13] S. LEFSCHETZ, *Introduction to Topology*, Princeton Univ. Press, Princeton, N.J., 1949.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

SYSTEM THEORY ON GROUP MANIFOLDS AND COSET SPACES*

R. W. BROCKETT†

Abstract. The purpose of this paper is to study questions regarding controllability, observability, and realization theory for a particular class of systems for which the state space is a differentiable manifold which is simultaneously a group or, more generally, a coset space. We show that it is possible to give rather explicit expressions for the reachable set and the set of indistinguishable states in the case of autonomous systems. We also establish a type of state space isomorphism theorem. These results parallel, and in part specialize to, results available for the familiar case described by $\dot{x}(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t)$. Our objective is to reduce all questions about the system to questions about Lie algebras generated from the coefficient matrices entering in the description of the system and in that way arrive at conditions which are easily visualized and tested.

1. Introduction. A standard assumption in modern control theory is that the state space is a vector space. This assumption is both valid and natural in many situations, but there is a significant class of problems for which it cannot be made. Typical of these are certain problems which arise in the control of the attitude of a rigid body. The state space in this case is not a vector space. Linearization often destroys the essence of the problem—even if one can work locally—and in any case new and different methods are needed for treating global questions.

In this paper we substitute the following hypothesis for the usual vector space assumptions. We let $\mathscr{F}$ and $\mathscr{C}$ be matrix groups and study

$$\dot{X}(t) = \left(A + \sum_{i=1}^{\nu} u_i(t)B_i\right)X(t), \qquad y(t) = \mathscr{C}X(t), \qquad X \in \mathscr{F},$$

where A and B_i belong to the Lie algebra associated with $\mathscr{F}$, the u_i are the controls, and the notation $\mathscr{C}X(t)$ is to be interpreted as being a coset in $\mathscr{F}$. We also study vector systems of a similar type in that we can view their evolution as occurring in a coset space. The results concern the explicit construction of the reachable set and a characterization of observability which is easily tested. Our main point is that this class of systems is in many ways not more difficult than linear systems of the usual type in $\mathbb{R}^n$.

There is a moderately large literature on the use of Chow's results [1] and related ideas to study controllability, including the work of Hermann, Kucera, Hermes, Haynes and Lobry (see [2]–[6]). This work is relevant here but we are directly interested in controllability only in so far as it contributes to the identification of a framework in which we can study a full range of system theoretic questions, including observability and realization theory. Notice that it is impossible to pass

* Received by the editors August 24, 1971, and in revised form October 27, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts 02138. This work was supported in part by the U.S. Office of Naval Research under the Joint Services Electronics Program by Contract N00014-67-A-0298-0006 and by the National Aeronautics and Space Administration under Grant NGR 22-007-172.

directly from controllability results to observability results in the present setup because there is no clear notion of duality. The main motivation for this work came from some work on Lie algebraic methods in differential equations (see [7]–[10]) and, above all, from being confronted with certain physical problems where linear theory was simply inadequate.

Some unpublished work [18], [19] by Jurdjevic and Sussmann is related to this paper. In particular they give in [19] an alternate proof of our Theorem 5 and make a serious study of the unsymmetric case (treated only superficially in Theorem 7 here). We also mention a recent paper by Elliott [20].

2. Examples. We postpone the development of the subject long enough to present a few simple examples which will help justify why the assumptions are set up the way they are.

Example 1 (Control systems design). Consider the problem of determining the gain, k, in the system

$$\dot{x}(t) = Ax(t) - k(t)bcx(t)$$

so as to achieve good performance relative to an index of the form

$$\eta = \int_0^\infty x'(t)Mx(t)\,dt, \qquad M = M' \geqq 0.$$

If a particular initial state is chosen and $k(\cdot)$ is selected so as to minimize η, then the performance might be bad relative to some other initial state. In cases where the initial state is not known, it is much more realistic to pick a collection of initial state vectors and to pick k in such a way as to minimize a weighted average of the individual performances. In fact, just to ensure stability it is necessary to average over at least n linearly independent initial states. If exactly n are chosen, then k should be regarded as controlling the evolution of the matrix equation

$$\dot{\Phi}(t) = (A - k(t)bc)\Phi(t), \qquad \Phi(0) = [x_1, x_2, \cdots, x_n].$$

The state space is then the space of nonsingular $n \times n$ matrices, $\mathscr{GL}(n)$.

Example 2 (Rigid body control). The orientation of a rigid body relative to some fixed set of axes is described by a 3×3 orthogonal matrix A which satisfies the differential equation

$$\begin{bmatrix} \dot{a}_{11}(t) & \dot{a}_{12}(t) & \dot{a}_{13}(t) \\ \dot{a}_{21}(t) & \dot{a}_{22}(t) & \dot{a}_{23}(t) \\ \dot{a}_{31}(t) & \dot{a}_{32}(t) & \dot{a}_{33}(t) \end{bmatrix} = \begin{bmatrix} 0 & \omega_3(t) & -\omega_2(t) \\ -\omega_3(t) & 0 & \omega_1(t) \\ \omega_2(t) & -\omega_1(t) & 0 \end{bmatrix} \begin{bmatrix} a_{11}(t) & a_{12}(t) & a_{13}(t) \\ a_{21}(t) & a_{22}(t) & a_{23}(t) \\ a_{31}(t) & a_{32}(t) & a_{33}(t) \end{bmatrix}.$$

The ω's themselves are usually controlled via the equations

$$\dot{\omega}_1(t) = [(I_2 - I_3)/I_1]\omega_2(t)\omega_3(t) + n_1(t)/I_1,$$
$$\dot{\omega}_2(t) = [(I_3 - I_1)/I_2]\omega_1(t)\omega_3(t) + n_2(t)/I_2,$$
$$\dot{\omega}_3(t) = [(I_1 - I_2)/I_3]\omega_1(t)\omega_2(t) + n_3(t)/I_3.$$

The state space for the first set of equations is $\mathscr{O}(3)$—the set of 3×3 orthogonal matrices. The state space for the second set of equations is $\mathbb{R}^3$—Cartesian 3-space.

For our present purpose suppose that the center of mass of the body is fixed and suppose that the observed output of this system is a pencil beam of light generated by a light source which is mounted in the body along a line passing through the center of mass. In this case the output is $\mathscr{C}X(t)$, where $\mathscr{C}$ is a subgroup which corresponds to a rotation about the pencil beam (an undetectable motion).

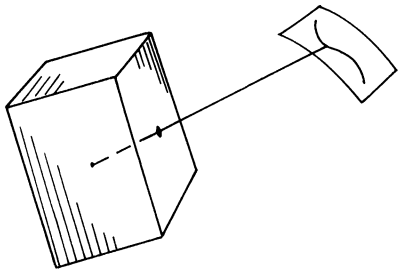


FIG. 1. *Illustration of the observability of a rigid body*

Example 3 (A model for DC to DC conversion). The electrical network shown in Fig. 1 contains switches which are to be manipulated in such a way as to transfer the energy stored on the capacitor 1 to capacitor 2. In order to have a sensible physical model we demand that there be exactly one path through the inductor at all times.

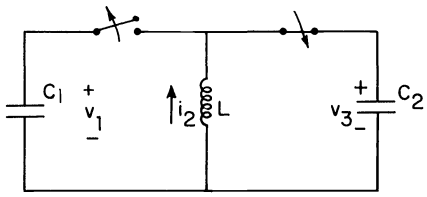


FIG. 2. *An electrical network for which energy is conserved*

The equations of motion are

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{bmatrix} = \begin{bmatrix} 0 & s_1(t) & 0 \\ -s_1(t) & 0 & s_2(t) \\ 0 & -s_2(t) & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix},$$

where $x_1 = v_1\sqrt{C_1}$, $x_3 = v_3\sqrt{C_2}$, $x_2 = i_2\sqrt{L}$ and s_1 and s_2 are dependent on the switch positions and take on the values 1 or 0. We have $s_1 = 1$ and $s_2 = 0$ if the switch on the left is closed, and we have $s_1 = 0$, $s_2 = 1$ if the switch on the right is closed.

3. Lie algebras and Lie groups. Let $\mathbb{R}^{n\times n}$ denote the set of real $n \times n$ matrices; $\mathbb{R}^{n\times n}$ is a vector space of dimension n^2. By a Lie algebra $\mathscr{L}$ in $\mathbb{R}^{n\times n}$ we understand a subset of $\mathbb{R}^{n\times n}$ which is a vector space and which has the property that if A and B belong to $\mathscr{L}$, thus so does $[A, B] = AB - BA$. If $\mathscr{L}_1$ and $\mathscr{L}_2$ are Lie algebras in $\mathbb{R}^{n\times n}$, and their intersection $\mathscr{L}_1 \cap \mathscr{L}_2$ is also a Lie algebra, then if A and B belong

to $\mathscr{L}_1$ and $\mathscr{L}_2$ and both are algebras, then $[A, B]$ belongs to both $\mathscr{L}_1$ and $\mathscr{L}_2$. The union $\mathscr{L}_1 \cup \mathscr{L}_2$ of two Lie algebras, the sum $\mathscr{L}_1 + \mathscr{L}_2$ of two Lie algebras and the commutator $[\mathscr{L}_1, \mathscr{L}_2]$ of two Lie algebras are not necessarily Lie algebras.

Given an arbitrary subset of $\mathbb{R}^{n\times n}$ we can add additional elements to it so as to imbed it in a Lie algebra. To obtain the smallest Lie algebra which contains a given set $\mathscr{N}$ we first add to $\mathscr{N}$ all linear combinations of elements in $\mathscr{N}$ so as to get a real vector space $\mathscr{N}_1$. Then we commute elements in $\mathscr{N}_1$ to get $\mathscr{N}_2 = \mathscr{N}_1 + [\mathscr{N}_1, \mathscr{N}_1]$; if this is not contained in $\mathscr{N}_1$, then we form $\mathscr{N}_3 = \mathscr{N}_2 + [\mathscr{N}_2, \mathscr{N}_2]$, etc. Clearly this process stops in a finite number of steps since at each stage we increase the dimension of the vector space by at least one and the dimension is upper bounded by n^2. We call this Lie algebra the Lie algebra generated by $\mathscr{N}$ and denote it by $\{\mathscr{N}\}_A$.

If $\mathscr{M}$ is a set of nonsingular matrices in $\mathbb{R}^{n\times n}$, we let $\{\mathscr{M}\}_G$ denote the multiplicative matrix group generated by $\mathscr{M}$, i.e., the smallest group in $\mathbb{R}^{n\times n}$ which contains $\mathscr{M}$ and is closed under multiplication and inversion. If $\mathscr{N}$ is a linear subspace of $\mathbb{R}^{n\times n}$, then the set

$$\mathscr{M} = \{M : M = e^{N_1} e^{N_2} \cdots e^{N_p},\ N_i \in \mathscr{N},\ p = 0, 1, 2, \cdots\}$$

contains no singular matrices since $\det(\exp N_i) = \exp(\operatorname{tr} N_i) > 0$. Clearly $\mathscr{M}$ is closed under multiplication and inversion and, in our notation,

$$\mathscr{M} = \{\exp \mathscr{N}\}_G.$$

Let $\mathscr{L}$ be a Lie algebra. At each point M in $\{\exp \mathscr{L}\}_G$ there is a one-to-one map ϕ_M from a neighborhood of 0 in $\mathscr{L}$ onto a neighborhood of M in $\{\exp \mathscr{L}\}_G$ which is defined by

$$\phi_M : \mathscr{L} \to \{\exp \mathscr{L}\}_G, \qquad \phi_M(L) = e^L M.$$

This map has a smooth inverse which shows that $\{\exp \mathscr{L}\}_G$ is a locally Euclidean space of dimension equal to the dimension of $\mathscr{L}$. We may check that the maps ϕ_M^{-1} satisfy the conditions for a C^∞-manifold in the sense of [11, p. 97]. Thus we may give $\{\exp \mathscr{L}\}_G$ the structure of a differentiable manifold. This justifies our referring to $\{\exp \mathscr{L}\}_G$ as a *group manifold.*

If $\mathscr{A}$ is a linear subspace of $\mathbb{R}^{n\times n}$ which is not necessarily a Lie algebra, we might inquire as to the relation between $\{\exp \mathscr{A}\}_G$ and $\{\exp \{\mathscr{A}\}_A\}_G$. Clearly the latter contains the former. The following theorem claims that they are identical.

THEOREM 1. *Let $\mathscr{A}_1, \mathscr{A}_2, \cdots, \mathscr{A}_p$ be a collection of linear subspaces of $\mathbb{R}^{n\times n}$. Then*

$$\{\exp \mathscr{A}_1, \exp \mathscr{A}_2, \cdots, \exp \mathscr{A}_p\}_G = \{\exp \{\mathscr{A}_1, \mathscr{A}_2, \cdots, \mathscr{A}_p\}_A\}_G.$$

Before proving Theorem 1 it is appropriate to make a few remarks about its relationship to the controllability literature. (Perhaps a glance at Theorem 5 would help at this point.) In considering the equation

$$\dot{X}(t) = \left(\sum_{i=1}^{m} u_i(t) A_i\right) X(t)$$

as a differential equation in $\mathscr{Gl}(n)$ it is clear from the theorem of Frobenius [12] that the solution passing through $X \in \mathscr{Gl}(n)$ lies in $\{\exp\{A_i\}_A\}_G X$ because at each point X in $\mathscr{Gl}(n)$. The collection $\{A_i X\}_A$ is an involutive distribution which spans the tangent space of $\{\exp\{A_i\}_A\}_G X$ at X. Wei and Norman [9] confirm this fact by giving (locally valid) formulas for the solution of this differential equation in terms of the functions $u_i(\cdot)$ and the structural constants of the Lie algebra generated by the A_i (without pointing out the differential geometric interpretation of the result). On the other hand, if we regard $\dot{X}(t) = (\sum_{i=1}^m u_i(t)A_i)X(t)$ as a control problem, then the natural question is not what manifold contains the solution, but rather what set can be attained from a given point, given freedom over the choice of $(u_1, u_2, \cdots, u_m)$. The results of Chow [1] (see also Hermann [2]) are applicable here. Chow showed under a suitable regularity condition that the set of points reachable for the vector system $\dot{x}(t) = \sum_{i=1}^m u_i(t) f_i[x(t)]$ using piecewise constant controls is the same as the set reachable for

$$\dot{x}(t) = \sum_{i=1}^{n} v_i(t) g_i[x(t)],$$

where $\{g_i(x)\}$ is a basis for the involutive distribution generated by $\{f_i(x)\}$. That is, $\{g_i(x)\}$ spans a vector space which includes $\{f_i(x)\}$ and is closed under the Lie bracket operation

$$[f, g] = \frac{\partial f}{\partial x} g - \frac{\partial g}{\partial x} f.$$

In our case the Lie bracket of $A_i X$ and $A_j X$ is $[A_i, A_j]X$. Thus we see that for the differential equation in question the reachable set includes $\{\exp\{A_i\}_A\}_G$, and the theorem of Frobenius ensures that it includes nothing more.

The proof of Theorem 1 given below could be shortened considerably by the use of these ideas. The reason for preferring the longer proof given here is that it is constructive, it is self-contained (nothing harder than the implicit function theorem is used) and it has the merit of proving a theorem about $n \times n$ matrices using the notation and tools natural to that subject.

Proof. We give a proof which relies on an implicit function theorem which, under suitable hypothesis, ensures the existence of a solution of α equations in $\beta < \alpha$ unknowns. (See [13, pp. 29–30].) We also need the Baker–Hausdorff formula which asserts that

$$e^{At} L e^{-At} = L + [At, L] + \frac{1}{2}[At, [At, L]] + \frac{1}{3!}[At, [At, [At, L]]] + \cdots.$$

Note that the norm of the $(n+1)$th term in this series is less than $\|L\| 2^n \|A\|^n / n!$ so that the series is majorized by the series $\|L\|[1 + 2\|A\| + (2\|A\|)^2/2! + \cdots] = \|L\| \cdot e^{2\|At\|}$ and hence is absolutely and uniformly convergent on $-T \leqq t\, T$ for all T.

Let $\{A_1, A_2, \cdots, A_n\} \subset \mathscr{A}_1 \cup \mathscr{A}_2 \cup \cdots \cup \mathscr{A}_p$ be a basis for $\mathscr{A}_1 + \mathscr{A}_2 + \cdots + \mathscr{A}_p$ and let $\mathscr{L}$ be the Lie algebra generated by $\{A_1, A_2, \cdots, A_r\}$. Assume

this algebra is of dimension q. There exists a basis for $\mathscr{L}$ which consists of terms of the form

$$
\begin{aligned}
&L_1 = A_1,\\
&L_2 = A_2,\\
&\cdots\\
&L_r = A_r,\\
&L_{r+1} = [A_{k(r+1)}, A_{l(r+1)}],\\
&L_{r+2} = [A_{k(r+2)}, A_{l(r+2)}],\\
&\cdots\\
&L_{r+s} = [A_{k(r+s)}, A_{l(r+s)}],\\
&L_{r+s+1} = [A_{k(r+s+1)}, [A_{l(r+s+1)}, A_{m(r+s+1)}]],\\
&\cdots\\
&L_q = [\cdots [A_{k(q)}, \cdots, A_{m(q)}] \cdots].
\end{aligned}
$$

We are quite explicit here because at certain points in our proof it is necessary to regard these expressions as formal expressions as opposed to matrices.

We introduce the following special notation. The operator EXP maps formal expressions into matrices. It is defined on A_i and its commutators (i.e., formal expressions such as A_i, $[A_i, A_j]$, $[A_i, [A_j, A_k]]$, etc.) as follows:

$$\mathrm{EXP}\, A_i t = e^{A_i t},$$

$$\mathrm{EXP}\,[A_i, A_j]t = \begin{cases} e^{A_i\sqrt{t}}\, e^{A_j\sqrt{t}}\, e^{-A_i\sqrt{t}}\, e^{-A_j\sqrt{t}}, & t \geqq 0,\\ e^{A_j\sqrt{|t|}}\, e^{A_i\sqrt{|t|}}\, e^{-A_j\sqrt{|t|}}\, e^{-A_i\sqrt{|t|}}, & t < 0. \end{cases}$$

The definition is completed by recursion. If A and B are formal expressions, then

$$\mathrm{EXP}\,[A, C]t = \begin{cases} (\mathrm{EXP}\, A\sqrt{t})(\mathrm{EXP}\, C\sqrt{t})(\mathrm{EXP}\, A\sqrt{t})^{-1}(\mathrm{EXP}\, C\sqrt{t})^{-1}, & t \geqq 0,\\ (\mathrm{EXP}\, C\sqrt{|t|})(\mathrm{EXP}\, A\sqrt{|t|})(\mathrm{EXP}\, C\sqrt{t})^{-1}(\mathrm{EXP}\, A\sqrt{t})^{-1}, & t < 0. \end{cases}$$

We now show that if B is a formal expression, then

$$\mathrm{EXP}\, Bt = I + Bt + o(t),$$

where $o(t)/t$ goes to zero as t goes to zero. We associate with each formal expression an integer n, called its degree, which is the largest number of times any particular entry of the formal expression is bracketed. To carry out this proof we use an induction on the degree of the expression.

Suppose A and C are two formal expressions of degree n or less. Suppose that we have shown that for any formal expression of degree n or less,

$$\mathrm{EXP}\, At = I + At + o(t).$$

Now in view of the definition of EXP we can expand EXP At and EXP Ct in a convergent power series involving fractional powers of t. Let us say

$$\text{EXP } At = I + At + E(t) + Dt^2 + o(t^2),$$

$$\text{EXP } Ct = I + C + G(t) + Ft^2 + o(t^2),$$

where E and G involve powers of t between t^1 and t^2 and E and F are the respective coefficients of t^2. The power series expansions for the inverses are then

$$(\text{EXP } At)^{-1} = I - At - E(t) + (A^2 - D)t^2 + o(t^2),$$

$$(\text{EXP } Ct)^{-1} = I - Ct - G(t) + (C^2 - F)t^2 + o(t^2),$$

as is verified by multiplication with the corresponding expressions for EXP At and EXP Ct respectively. Now using these expressions we have for t nonnegative,

$$\begin{aligned}
&\text{EXP } Bt = \text{EXP}\,[A, C]t \\
&\quad = (I + A\sqrt{t} + E(t) + Dt^2 + o(t))(I + C\sqrt{t} + G(\sqrt{t}) + Ft + o(t)) \\
&\qquad \cdot (I - A\sqrt{t} + E(t) + Dt^2 + o(t))((I - C\sqrt{t} - G(\sqrt{t}) + C^2 - F)t + o(t)) \\
&\quad = I + [A, C]t + Ft + A^2t/2 + A^2t/2 - A^2t + (C^2 - F)t - C^2t + o(t) \\
&\quad = I + [A, C]t + o(t),
\end{aligned}$$

and the case $t < 0$ leads to the same result.

Well-known properties of the matrix exponential function let one conclude that for $|t| > 0$, EXP Bt is continuously differentiable. The above argument shows that EXP Bt is differentiable with respect to t in a neighborhood of $t = 0$ and

$$\frac{d}{dt}\,\text{EXP } Bt|_{t=0} = B.$$

Hence we have for the basis elements L_ν,

$$\frac{d}{dt}\,\text{EXP } L_\nu t|_{t=0} = L_\nu.$$

Now consider a function of $u = (u_1, u_2, \cdots, u_q)$ and $v = (v_1, v_2, \cdots, v_q)$ which maps $\mathbb{R}^q \times \mathbb{R}^q$ into $\mathbb{R}^{n\times n}$ and which is defined by

$$\begin{aligned}
F(u, v) = &-I + (\text{EXP } L_1u_1)(\text{EXP } L_2u_2)\cdots(\text{EXP } L_qu_q) \\
&\cdot \exp(-L_1v_1 - L_2v_2 \cdots - L_qv_q).
\end{aligned}$$

Clearly $F(0, 0) = 0$. Now the linear approximation of F at $(u, v) = (0, 0)$ is given by

$$\begin{aligned}
F_{(u,v)}(u, v)|_{(0,0)}(\delta u, \delta v) = &L_1\delta u_1 + L_2\delta u_2 + \cdots + L_q\delta u_q - L_1\delta v_1 \\
&- L_2\delta v_2 - \cdots - L_q\delta v_q
\end{aligned}$$

so that the range space of $F_{(u,v)}(u, v)|_{(0,0)}(u, 0)$ is the q-dimensional subspace of

$\mathbb{R}^{n\times n}$ spanned by $\{L_i\}$. Now $F(u, v) + I$ is a finite product of exponentials which we write as

$$F(u,v) + I = \exp(A_{i_1}u_{i_1}^{p_1}\exp(A_{i_2}u_{i_2}^{p_2})\cdots\exp(A_{i_\nu}u_{i_\nu}^{p_\nu}) \cdot \exp(-L_1v_1 - L_2v_2 - \cdots - L_qv_q).$$

Since the Baker–Hausdorff formula lets one write

$$e^{A_it}A_k\,e^{-A_it} = A_k + [A_i, A_k]t + \cdots,$$

we see that $e^{A_it}A_k\,e^{-A_it}$ belongs to the Lie algebra generated by the A's. Moreover, it is continuous with respect to t and at $t = 0$ takes on the value A_k. Using this result repeatedly we see that for each $\{u_{i_k}\}$ we can find R_{i_k} in $\mathscr{L}$ such that

$$\exp(A_{i_1}u_{i_1}^{p_1})\exp(A_{i_2}u_{i_2}^{p_2})\cdots\exp(A_{i_k}u_{i_k}^{p_k})A_{i_k}\exp(A_{i_{k+1}}u_{i_{k+1}}^{p_{k+1}})\cdots\exp(A_{i_\nu}u_{i_\nu}^{p_\nu}) \cdot \exp(L_1v_1 + L_2v_2 + \cdots + L_qv_q) = (F(u,v) + I)R_{i_k}(u_{i_1}, u_{i_2}, \cdots, u_{i_\nu})$$

simply by pushing A_k past the exponentials one at a time. Clearly $R_{i_k}(0, 0, \cdots) = A_{i_k}$. Thus we see that for a and b small,

$$F_{(u,v)}(u,v)|_{(a,b)}(\delta u, \delta v) = (F(a,b) + I)\left(\sum_{i=1}^{q} S_i(a,b)\delta u_i + H_i(a,b)\delta v_i\right)$$

for some $S_i(a, b)$ and $H_i(a, b)$ in $\mathscr{L}$. Since $S_i(0, 0) = H_i(0, 0) = L_i$ and since S_i and H_i depend continuously on their arguments, this establishes that the Jacobian of the map $F: \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}^{n\times n}$ must have rank q in a neighborhood of $(0, 0)$; and hence, by the implicit function theorem cited earlier, there exists an $\varepsilon > 0$ and a map $\phi: \mathbb{R}^q \to \mathbb{R}^q$ such that if $\|v\| < \varepsilon$, then

$$F(\phi(v), v) = 0.$$

Since $F(u, v) = 0$ implies that

$$\text{EXP}\, L_1u_1\, \text{EXP}\, L_2u_2 \cdots \text{EXP}\, L_qu_q = \exp(L_1v_1 + L_2v_2 + \cdots + L_qv_q)$$

we conclude that there exists $\varepsilon_1 > 0$ such that if $L \in \mathscr{L}$ and $\|L\| < \varepsilon_1$, we can write

$$e^L = \exp(A_{i_1}u_{i_1})\exp(A_{i_2}u_{i_2})\cdots\exp(A_{i_\nu}u_{i_\nu}).$$

Now for any $L \in \mathscr{L}$ it follows that $\|(1/m)L\| < \varepsilon_1$ for some integer m, and thus we can express e^L as $e^{L/m}\,e^{L/m}\cdots e^{L/m}$. Likewise we can express $e^{L_1}\,e^{L_2}\cdots e^{L_q}$ in this form.

Let $\mathscr{K}$ and $\mathscr{L}$ be Lie algebras in $\mathbb{R}^{n\times n}$. It can happen that $\{\exp\mathscr{K}\}_G$ is a bounded subset of $\mathbb{R}^{n\times n}$ which is not closed, and it can happen that the closure of $\{\exp\mathscr{K}\}_G$ equals $\{\exp\mathscr{L}\}_G$ with $\mathscr{K} \neq \mathscr{L}$. The skew-line on the torus [11] is an easy example. Also, $\{\exp\mathscr{K}\}_G$ is not necessarily simply connected. Nonetheless, we have the following result which we deduce from Theorem 1 rather then sending the reader to the literature.

COROLLARY 1. *If $\mathscr{K}$ and $\mathscr{L}$ are Lie algebras in $\mathbb{R}^{n\times n}$, then $\{\exp\mathscr{K}\}_G \subset \{\exp\mathscr{L}\}_G$ if and only if $\mathscr{K} \subset \mathscr{L}$, and $\{\exp\mathscr{K}\}_G = \{\exp\mathscr{L}\}_G$ if and only if $\mathscr{K} = \mathscr{L}$.*

Proof. For both statements the sufficiency is obvious. To establish necessity in the first case notice that if $\{\exp \mathscr{K}\}_G \subset \{\exp \mathscr{L}\}_G$, then by Theorem 1,

$$\{\exp \mathscr{K}\}_G = \{\exp \mathscr{K}, \exp \mathscr{L}\}_G = \{\exp \{\mathscr{K}, \mathscr{L}\}_A\}_G.$$

Suppose $\mathscr{L}$ is of dimension n. To obtain a proof by contradiction, suppose that $\mathscr{K}$ is not contained in $\mathscr{L}$. Then $\{\mathscr{K}, \mathscr{L}\}_A$ is of dimension $n+1$ or greater. Then $\exp \mathscr{L}$ is an n-dimensional manifold and $\{\exp \{\mathscr{K}, \mathscr{L}\}\}$ is not, which contradicts $\{\exp \mathscr{L}\} = \{\exp \{\mathscr{L}, \mathscr{K}\}\}$. To establish necessity in the second case repeat this argument verbatim but with "contained in" replaced by "equals" both verbally and symbolically.

The notation $\mathrm{ad}_A^0 B = B, \mathrm{ad}_A^1 B = [A, B], \mathrm{ad}_A^2 B = [A, [A, B]]$, etc., is standard. If $\mathscr{K}$ and $\mathscr{L}$ are Lie algebras, we use the notation $\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A$ to denote the Lie algebra generated by $\mathscr{K}$ under commutation with elements of $\mathscr{L}$. That is,

$$\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A = \{\mathscr{K}, [\mathscr{L}, \mathscr{K}], \cdots, [\mathscr{L}, [\mathscr{L} \cdots [\mathscr{L}, \mathscr{K}] \cdots]] \cdots\}_A.$$

This algebra may also be described as the intersection of all Lie algebras which contain $\mathscr{K}$ and are closed under commutation with $\mathscr{L}$. If $\mathscr{F}$ and $\mathscr{G}$ are groups, we introduce an analogous notation. The smallest group which contains $\mathscr{F}$ and all products of the type GFG^{-1} for G in $\mathscr{G}$ and F in $\mathscr{F}$ will be denoted by $\{\mathrm{AD}_{\mathscr{G}} \mathscr{F}\}_G$. This group may be described as the intersection of all groups which contain $\mathscr{F}$ and are closed under conjugation with elements of $\mathscr{G}$. If $\mathscr{G}$ is $\{\exp \mathscr{L}\}_G$ and $\mathscr{F}$ is $\{\exp \mathscr{K}\}_G$, then clearly $\{\mathrm{AD}_{\mathscr{G}} \mathscr{F}\}_G$ consists of products of terms of the form

$$M = e^{L_1} e^{K_1} e^{-L_1} e^{L_2} e^{K_2} e^{-L_2} \cdots e^{L_m} e^{K_m} e^{-L_m}.$$

THEOREM 2. *Let $\mathscr{K}$ and $\mathscr{L}$ be Lie algebras in $\mathbb{R}^{n \times n}$. Then*

$$\left\{ \bigcup_{M \in \{\exp \mathscr{L}\}_G} M \mathscr{K} M^{-1} \right\}_A = \{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A$$

and

$$\{\mathrm{AD}_{\{\exp \mathscr{L}\}_G} \{\exp \mathscr{K}\}_G\}_G = \{\exp \{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A\}_G.$$

Proof. From the Baker–Hausdorff formula we see at once that if L belongs to $\mathscr{L}$ and K belongs to $\mathscr{K}$, then $e^L K e^{-L}$ belongs to $\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A$. Thus the right side of the first equality in question contains the left. On the other hand, expressions of the following type belong to the left side:

$$e^{\alpha L} \frac{1}{\alpha} K e^{-\alpha L} - \frac{1}{\alpha} K = [L, K] + o(\alpha),$$

$$e^{\alpha L} \left(\frac{1}{\alpha} [K, L] + o(\alpha) \right) e^{-\alpha L} - \frac{1}{\alpha} [K, L] + o(\alpha) = [L[L, K]] + o(\alpha), \quad \text{etc.}$$

Since $\mathbb{R}^{n \times n}$ is a finite-dimensional space and since $\{\bigcup_{M \in \{\exp \mathscr{L}\}_G} M \mathscr{K} M^{-1}\}_A$ is a linear subspace, it is closed. Thus $[L, K]$, $[L[L, K]] \cdots$ belong to this set and the first equality is seen to hold.

The second statement is obtained by exponentiating the first. This gives

$$\left\{ \exp \left\{ \bigcup_{M \in \{\exp \mathscr{L}\}_G} M \mathscr{L} M^{-1} \right\}_A \right\}_G = \{\exp \{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A\}_G,$$

but since $e^L e^K e^{-L} = \exp(e^L K e^{-L})$, we see that

$$\left\{\exp\left\{\bigcup_{M\in\{\exp\mathscr{L}\}G} M\mathscr{L}M^{-1}\right\}_A\right\}_G = \{\mathrm{AD}_{\{\exp\mathscr{L}\}_G}\{\exp\mathscr{L}\}_G\}_G$$

so the result follows.

The next theorem states a purely group theoretic result which, although easily proved, is stated formally because we need it in our study of observability.

THEOREM 3. *Let $\mathscr{H}$ and $\mathscr{R}$ be subgroups of a group $\mathscr{G}$. Let $\mathscr{P}$ be the subset of $\mathscr{G}$ defined as*

$$\mathscr{P} = \{P: RPR^{-1} \in \mathscr{H}, \text{ all } R \in \mathscr{R}\}.$$

Then $\mathscr{P}$ is a subgroup of $\mathscr{H}$, $\mathscr{R}\mathscr{P}$ is a subgroup of $\mathscr{G}$, and $\mathscr{P}$ is a normal subgroup of $\mathscr{R}\mathscr{P}$. Thus $\mathscr{R} \cap \mathscr{P}$ is a normal subgroup of $\mathscr{R}$ and $\mathscr{R}\mathscr{P}/\mathscr{P}$ is isomorphic to $\mathscr{R}/\mathscr{R} \cap \mathscr{P}$.

Proof. Suppose P_1 and P_2 belong to $\mathscr{P}$. Then for each R in $\mathscr{R}$ there exist $H_1(R)$ and $H_2(R)$ in $\mathscr{H}$ such that $RP_1R^{-1}(RP_2^{-1}R^{-1}) = H_1(R)\cdot[H_2(R_2)]^{-1}$. Since $\mathscr{H}$ is a group, this means $RP_1P_2^{-1}R^{-1}$ belongs to $\mathscr{H}$, and thus that $\mathscr{P}$ is a subgroup of $\mathscr{G}$. Clearly it is a subgroup of $\mathscr{H}$ since the choice $R = I$ is possible. To see that $\mathscr{R}\mathscr{P}$ is a group, note that if R_1 and R_2 belong to $\mathscr{R}$ and P_1 and P_2 belong to $\mathscr{P}$, then

$$R_1P_1(R_2P_2)^{-1} = R_1P_1R_2^{-1}R_2P_2^{-1}R_2^{-1} = R_1R_2^{-1}(R_2P_1R_2^{-1})(R_2P_2^{-1}R_2^{-1}).$$

Since $\mathscr{R}$ is a group and since $\mathscr{P}$ is a group which has the property that if P belongs to $\mathscr{P}$ then so does RPR^{-1} for each R in $\mathscr{R}$, we see that this product belongs to $\mathscr{R}\mathscr{P}$. Clearly $\mathscr{P}$ is a normal subgroup of $\mathscr{R}\mathscr{P}$ since $RP\mathscr{P}P^{-1}R^{-1} = \mathscr{P}$ for each RP in $\mathscr{R}\mathscr{P}$. By the second isomorphism theorem (Rotman [14, p. 26]) $\mathscr{R} \cap \mathscr{P}$ is normal in $\mathscr{R}$ and $\mathscr{R}\mathscr{P}/\mathscr{P} \simeq \mathscr{R}/\mathscr{R} \cap \mathscr{P}$.

We now state and prove a Lie algebraic analogue of this theorem. Algebraic tests for observability will be derived from this result.

THEOREM 4. *Let $\mathscr{H}$ and $\mathscr{L}$ be Lie algebras in $\mathbb{R}^{n\times n}$. Let $\mathscr{P}$ be defined as*

$$\mathscr{P} = \{P: RPR^{-1} \in \{\exp\mathscr{H}\}_G, \text{ all } R \in \{\exp\mathscr{L}\}_G\}.$$

If $\mathscr{K}$ is a Lie algebra in $\mathbb{R}^{n\times n}$, then $\{\exp\mathscr{K}\}_G \subset \mathscr{P}$ if and only if $\{\mathrm{ad}_{\mathscr{L}}\mathscr{K}\}_A \subset \mathscr{H}$. There exists a unique Lie algebra $\mathscr{K}_1$ such that $\{\mathrm{ad}_{\mathscr{L}}\mathscr{K}_1\}_A \subset \mathscr{H}$ and $\mathscr{K}_1$ contains all other Lie algebras having this property.

Proof. Suppose $\{\mathrm{ad}_{\mathscr{L}}\mathscr{K}\}_A \subset \mathscr{H}$. Then for L_i in $\mathscr{L}$ and K_i in $\mathscr{K}$ we see from Theorem 2 that $\{\exp\mathscr{H}\}_G$ contains

$$RPR^{-1} = e^{L_1}e^{L_2}\cdots e^{L_p}e^{K_1}e^{K_2}\cdots e^{K_q}e^{-L_p}\cdots e^{-L_2}e^{-L_1}.$$

By the hypothesis and Corollary 1, $\{\exp\{\mathrm{ad}_{\mathscr{L}}\mathscr{K}\}_A\}_G \subset \{\exp\mathscr{H}\}_G$.

On the other hand, if for all L_i in $\mathscr{L}$ and all K_i in $\mathscr{K}$ we have

$$RPR^{-1} = e^{L_1}e^{L_2}\cdots e^{L_p}e^{K_1}e^{K_2}\cdots e^{K_q}e^{-L_p}\cdots e^{-L_2}e^{-L_1} \in \{\exp\mathscr{H}\}_G,$$

then since $\{\exp\mathscr{H}\}_G$ is a group, we see that $\{\mathrm{AD}_{\{\exp\mathscr{L}\}_G}\{\exp\mathscr{K}\}_G\}_G \subset \{\exp\mathscr{H}\}_G$, and again from Theorem 2 and Corollary 1 we see that $\{\mathrm{ad}_{\mathscr{L}}\mathscr{K}\}_A \subset \mathscr{H}$.

Finally, notice that if $\{\mathrm{ad}_{\mathscr{L}}\mathscr{K}_1\}_A \subset \mathscr{H}$ and $\{\mathrm{ad}_{\mathscr{L}}\mathscr{K}_2\}_A \subset \mathscr{H}$, then $\{\mathrm{ad}_{\mathscr{L}}(\mathscr{K}_1 + \mathscr{K}_2)\}_A \subset \mathscr{H}$. Thus there is a largest Lie algebra with this property.

4. Controllability on group manifolds. The first question of a system theoretic character which we investigate is that of controllability. Since we want to emphasize global results, we work with the most elementary type of evolution equation appropriate to our present setting, namely,

$$\dot{X}(t) = (A + \sum u_i(t)B_i)X(t).$$

The choice of control affects the direction in which X moves. However, A is a constant over which there is no control. This evolution equation has the property that the change of variable $X \to XP$ for P nonsingular leaves the equation unchanged. This invariance gives the vector field which a given choice of $\{u_i(t)\}$ establishes on $\mathscr{Gl}(n)$, a particularly simple form.

THEOREM 5. *Consider the linear dynamical system*

$$\dot{X}(t) = \left(\sum_{i=1}^{m} u_i(t)B_i\right)X(t), \qquad X = \text{an } n \times n \text{ matrix}.$$

Given a time $t_a > 0$ and given two nonsingular matrices X_1 and X_2, there exist piecewise continuous controls which steer the state from X_1 at $t = 0$ to X_2 at $t = t_a$ if and only if $X_2X_1^{-1}$ belongs to $\{\exp\{B_i\}_A\}_G$.

Proof. Sufficiency. Theorem 1 asserts that any matrix M in $\{\exp\{B_i\}_A\}_G$ can be written as a finite product, say

$$M = e^{B_{i_1}\alpha_1}\, e^{B_{i_2}\alpha_2} \cdots e^{B_{i_m}\alpha_m}.$$

Suppose $X_2X_1^{-1} = M$. Divide the interval $0 \leqq t \leqq t_a$ up into m equal intervals $[t_i, t_{i+1})$ whereby $t_i = i \cdot t_a/m$. Let $t_a/m = \beta^{-1}$. On the interval $[0, t_1)$ all controls are zero except the i_mth control, which takes on the value $\alpha_m\beta$. On the interval $[t_1, t_2)$ all controls are zero except the i_{m-1}th, which takes on the value $\alpha_{m-1}\beta$, etc., down to the last interval on which all controls are zero except the i_1st, which takes on the value $\alpha_1\beta$. Since the differential equation is linear and constant on each of the subintervals, the solution is a product of exponentials and the result follows.

Necessity. To show that X_2 cannot be reached from X_1 unless $X_2X_1^{-1}$ is of the form $e^{L_m}\, e^{L_{m-1}} \cdots e^{L_1}$ we assume the contrary and obtain a contradiction. Suppose that $u_1(\cdot), \cdots, u_m(\cdot)$ is a control which steers the system from X_1 at $t = 0$ to X_2 at $t = t_a$. By Theorem 1 of [9] we know that there exists a sequence of times $t_0, t_1, t_2, \cdots, t_m$ such that on each of the subintervals $[t_i, t_{i+1}]$ the transition matrix of

$$\dot{X}(t) = \left(\sum_{i=1}^{q} u_i(t)B_i\right)x(t)$$

can be written as $e^{H_i(t)}$ for some $H_i(\cdot)$ in $\mathscr{L}$. Thus we can write

$$X(t_a) = e^{L_m}\, e^{L_{m-1}} \cdots e^{L_1}X_0,$$

which establishes the contradiction.

As an application which emphasizes the ease with which we can study global questions using this theorem we observe the following results relating to the classical groups. Here J is given by

$$J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix},$$

and a matrix is called symplectic if $\theta' J\theta = J$.

THEOREM 6. *Consider the system of Theorem* 5. *Given a time* $t_a > 0$ *and given two nonsingular* $n \times n$ *matrices* X_1 *and* X_2 *with* $\det X_1 X_2 > 0$, *there exists a piecewise continuous control which steers the state from* X_1 *at* $t = 0$ *to* X_2 *at* $t = t_a$ *if* $\{B_i\}_A$:

(i) *spans* $\mathbb{R}^{n\times n}$,

(ii) *spans the* $(n^2 - 1)$*-dimensional subspace of* $\mathbb{R}^{n\times n}$ *consisting of the zero trace matrices and* $\det X_1 = \det X_2$,

(iii) *spans the* $(n(n+1)/2)$*-dimensional subset of* $\mathbb{R}^{n\times n}$ *consisting of the set of matrices which satisfy* $JA + A'J = 0$ *and* $X_2 X_1^{-1}$ *is symplectic,*

(iv) *spans the* $(n(n-1)/2)$*-dimensional subset of* $\mathbb{R}^{n\times n}$ *consisting of all skew symmetric matrices, and* $X_2 X_1^{-1}$ *is orthogonal.*

Proof. As is well known, any nonsingular matrix can be written as θR with $\theta'\theta = I$ and $R = R' > 0$. Also, real orthogonal matrices with positive determinants and real symmetric positive definite matrices have real logarithms. Moreover, in case (iii) the factors in the polar representation inherit the property of the group itself, which is to say that the θ and R in the polar representation of a symplectic matrix are symplectic. To complete the proof we need only invoke Theorem 5 since the previous remarks justify our writing $X_2 X^{-1} = e^{\Omega} e^{S}$ with $\Omega = -\Omega'$ and $S = S'$ both in the appropriate Lie algebras.

The results of Theorems 5 and 6 are somewhat unsatisfactory in that the A term is absent. The following theorem describes one way in which this can be relaxed.

THEOREM 7. *Consider the linear dynamical system*

$$\dot{X}(t) = \left(A + \sum_{i=1}^{v} u_i(t) B_i\right) X(t), \qquad X = \text{an } n \times n \text{ matrix}.$$

Suppose that $[\mathrm{ad}_A^k B_i, B_j] = 0$ *for* $i, j = 1, 2, \cdots, v$ *and* $k = 0, 1, \cdots, n^2 - 1$. *Let* $\mathscr{H}$ *be the linear subspace of* $\mathbb{R}^{n\times n}$ *spanned by* $\mathrm{ad}_A^k B_i$ *for* $i = 1, 2, \cdots, v$ *and* $k = 0, 1, \cdots, n^2 - 1$. *Then given a time* $t_a > 0$ *and two* $n \times n$ *matrices* X_1 *and* X_2 *there exist continuous controls which steer the system from the state* X_1 *at* $t = 0$ *to the state* X_2 *at* $t = t_a$ *if and only if there exists* H *in* $\mathscr{H}$ *such that*

$$X_2 = e^{At_a} e^{H} X_1.$$

Proof. First of all, notice that

$$\begin{aligned}
\frac{d^k}{dt^k}[e^{At} B_i e^{-At}, B_j]\Big|_{t=0} &= \frac{d^{k-1}}{dt^{k-1}}[e^{At}[A, B_i] e^{-At}, B_j]\Big|_{t=0} \\
&= \frac{d^{k-2}}{dt^{k-2}}[e^{At}(\mathrm{ad}_A^2 B_i) e^{-At}, B_j]\Big|_{t=0} \\
&\quad \cdots \\
&= [e^{At}(\mathrm{ad}_A^k B_i) e^{-At}, B_j]|_{t=0} \\
&= [\mathrm{ad}_A^k B_i, B_j].
\end{aligned}$$

Thus $[e^{At} B_i e^{-At}, B_j]$ is identically zero if $[\mathrm{ad}_A^k B_i, B_j] = 0$ for $k = 0, 1, 2, \cdots$. However, ad_A is a linear operator from an n^2-dimensional space into itself so that by the Cayley–Hamilton theorem all powers above $n^2 - 1$ are lineary dependent

on the first $n^2 - 1$. Thus under the hypothesis of the theorem statement, $[e^{At}B_i e^{-At}, B_j]$ vanishes identically. Also

$$\begin{aligned} 0 &= e^{At}B_i e^{-At}B_j - B_j e^{At}B_i e^{-At} \\ &= e^{A\sigma}(e^{At}B_i e^{-At})B_j e^{-A\sigma} - e^{A\sigma}B_j e^{At}B_i e^{-At} e^{-A\sigma}. \end{aligned}$$

Now let $t + \sigma = \beta$ and $\gamma = \sigma$. Thus for all β and γ,

$$0 = [e^{A\beta}B_i e^{-A\beta}, e^{A\gamma}B_j e^{-A\gamma}].$$

For the purpose of solving the differential equation we introduce $Z(t) = e^{-At}X(t)$ and observe that

$$\dot{Z}(t) = \left(\sum_{i=1}^{v} u_i(t)\, e^{-At}B_i\, e^{At}\right) Z(t).$$

But recall (see e.g. Martin [15]) that the solution of $\dot{Z}(t) = B(t)Z(t)$ is $\exp \int_0^t B(\sigma)\, d\sigma$ if $[B(t), B(\sigma)]$ vanishes for all t and σ. Thus we can write

$$Z(t) = \exp\left(\int_0^t \sum_{i=1}^{n} u_i(t)\, e^{-At}B_i\, e^{At}\, dt\right) Z(0).$$

It is a well-known and frequently used fact (e.g., [16, p. 79]) that the image space of the map taking continuous functions into $\mathbb{R}^p$ according to the rule $x = L(u) = \int_0^{t_1} e^{A\sigma}bu(\sigma)\, d\sigma$, is spanned by the first p derivatives of $e^{At}b$ evaluated at zero. Using this fact here we see that for each H in $\mathscr{H}$ and $t_a > 0$ we have a continuous u defined on $[0, t_a]$ such that

$$Z(t_a) = e^{H}Z(0).$$

Therefore in terms of X we see that we can reach at t_a any X which can be expressed as $e^{At_a}\, e^{H}x(0)$ with H in $\mathscr{H}$.

As an application of this result we derive a familiar relationship.

Example 4. Consider the system in $\mathbb{R}^n$,

$$\dot{x}(t) = \tilde{A}x(t) + \sum_{i=1}^{m} b_i u_i(t); \qquad x(0) \text{ is given}.$$

Related to this is the matrix system in $\mathbb{R}^{(n+1)\times(n+1)}$:

$$\dot{X}(t) = \begin{bmatrix} \tilde{A} & 0 \\ 0 & 0 \end{bmatrix} X(t) + \sum_{i=1}^{m} u_i(t) \begin{bmatrix} 0 & b_i \\ 0 & 0 \end{bmatrix} X(t).$$

Let A and B_i be the matrices appearing in this expression. In this case $[\mathrm{ad}_A^k B_i, B_j]$ vanishes as required and so the reachable set from $x(0) = I$ is

$$\mathscr{R}(t) = \exp \begin{bmatrix} At & 0 \\ 0 & 0 \end{bmatrix} \{\exp H;\, H \in \mathscr{H}\},$$

where $\mathscr{H}$ is the subspace spanned by $\mathrm{ad}_A^k B_i$. A computation gives

$$\mathrm{Ad}_A^k B_i = \begin{bmatrix} 0 & A^k b_i \\ 0 & 0 \end{bmatrix}$$

so that the reachable set at t is

$$\mathscr{R} = \left\{ X : X = \begin{bmatrix} e^{At} & H \\ 0 & 1 \end{bmatrix}; H \in \text{range}\,(B, AB, \cdots, A^{n-1}B) \right\},$$

where we have used the fact that $e^{At}\mathscr{K} = \mathscr{K}$ for all t and $\mathscr{K} = \text{range}\,(B, AB, \cdots, A^{n-1}B)$.

5. Observability. In order to get a theory having a scope comparable to linear theory, it is necessary to treat observability. The choice of an appropriate form of the observational equation is critical for the success of the overall theory. As it turns out, the natural choice is indicated by the second example in § 2.

Let $\mathscr{F}$ be a matrix group and let $\mathscr{C}$ be a subgroup. Consider the system evolving in $\mathscr{F}$,

$$\dot{X}(t) = \left(A + \sum_{i=1}^{v} u_i(t) B_i \right) X(t), \qquad y(t) = \mathscr{C} X(t),$$

by which we mean that instead of observing $X(t)$ directly, we observe to what equivalence class $X(t)$ belongs with respect to the equivalence relation in $\mathscr{F}$ defined by $\mathscr{C}$. Thus $y(t)$ takes on values in the coset space $\mathscr{F}/\mathscr{C}$ which is generally not a group manifold (see § 7).

We call two states X_1 and X_2 *distinguishable* if there exists some control which gives rise to different outputs for the two starting states. In general the zero control is not adequate to distinguish between all states which are distinguishable as contrasted with the situation one finds for linear systems.

THEOREM 8. *Let $\mathscr{C}$ be a matrix group and suppose that the set of points $\mathscr{R}$ reachable from the identity for the system*

$$\dot{X}(t) = \left(A + \sum_{i=1}^{v} u_i(t) B_i \right) X(t), \qquad y(t) = \mathscr{C} X(t),$$

is a group. Then the set of initial states which are indistinguishable from the identity is given by

$$\mathscr{P} = \{ P : RPR^{-1} \in \mathscr{C} \text{ for all } R \in \mathscr{R} \}.$$

$\mathscr{P}$ is a normal subgroup of $\mathscr{P}\mathscr{R}$ and a subgroup of $\mathscr{C}$.

Proof. Suppose that X is a starting state for the given equation which is indistinguishable from the identity. That means that for each R in $\mathscr{R}$ there is $C(R)$ in $\mathscr{C}$ such that

$$C(R)RX = R.$$

Since $\mathscr{R}$ and $\mathscr{C}$ are groups, we can take inverses to get

$$RXR^{-1} \in \mathscr{C}.$$

Thus the set $\mathscr{P}$ is exactly those states indistinguishable from the identity. The remainder of the conclusions come from Theorem 3.

THEOREM 9. *Let $\mathscr{H}$ and $\mathscr{L}$ be Lie algebras in $\mathbb{R}^{n\times n}$, and suppose that all the points reachable from the identity for*

$$\dot{X}(t) = \left(A + \sum_{i=1}^{v} u_i(t) B_i \right) X(t), \qquad y(t) = \{\exp \mathscr{H}\}_G X(t),$$

are $\{\exp \mathscr{L}\}_G$. *Then the set of initial states* $\mathscr{P}$ *which are indistinguishable from the identity contains* $\{\exp \mathscr{K}\}_G$ *if and only if* $\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A \subset \mathscr{K}$. *Therefore a necessary condition for all states to be distinguishable from the identity is that* $\mathscr{K}$ *contain no subalgebra* $\mathscr{K}$ *such that* $\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A \subset \mathscr{K}$.

Proof. Theorem 8 gives a characterization of $\mathscr{P}$ which permits one to bring to bear Theorem 4. Theorem 4 immediately gives the desired result.

One might be tempted to conclude that if there is no nontrivial algebra $\mathscr{K}$ meeting the requirements of Theorem 9, then all initial states are distinguishable. This is not true because $\mathscr{P}$ can be a discrete subgroup and hence not trivial and yet not expressible as $\{\exp \mathscr{K}\}_G$ for any Lie algebra $\mathscr{K}$. The next example illustrates this.

Example 5. In the numerical integration of the equations of motion of a rigid body one usually avoids Euler angle representation and uses instead quaternion or direction cosine representations. As is well known, the group of unit quaternions covers $\mathscr{SO}(3)$ twice. This causes an ambiguity in going from $\mathscr{SO}(3)$ to the group of unit quaternions. This example illustrates this idea. Consider an equation in the group of unit quaternions $\mathscr{Q}$ which we parametrize in the usual way ($a^2 + b^2 + c^2 + d^2 = 1$):

$$\frac{d}{dt}\begin{bmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{bmatrix} = \begin{bmatrix} 0 & u_1 & u_2 & u_3 \\ -u_1 & 0 & u_3 & -u_2 \\ -u_2 & -u_3 & 0 & u_1 \\ -u_3 & +u_2 & -u_1 & 0 \end{bmatrix}\begin{bmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{bmatrix},$$

$$y(t) = \mathscr{C}\begin{bmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{bmatrix},$$

where $\mathscr{C}$ is the subgroup given by

$$\mathscr{C} = \{\exp \mathscr{H}\}_G, \qquad \mathscr{H} = \begin{bmatrix} 0 & \alpha & 0 & 0 \\ -\alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha \\ 0 & 0 & -\alpha & 0 \end{bmatrix}.$$

Now it is true that $\{\exp \mathscr{H}\}_G$ includes I and $-I$, and it is also true that this pair of elements forms a normal subgroup of $\mathscr{Q}$. Thus I as an initial state cannot be distinguished from $-I$. Yet there is no nontrivial Lie algebra $\mathscr{K}$ such that $\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A \subset \mathscr{K}$.

6. Realization theory. One of the central results in linear system theory is the fact that any two time-invariant, controllable and observable realizations of a given time-invariant input–output map are related to each other in a very simple way. Our purpose here is to establish a similar theorem in this context.

Suppose we have two systems

$$\dot{X}(t) = \left(\sum_{i=1}^{m} u_i(t)B_i\right)X(t), \qquad y(t) = \mathscr{C}X(t),$$

$$\dot{Z}(t) = \left(\sum_{i=1}^{m} u_i(t)G_i\right)Z(t), \qquad y(t) = \mathscr{H}Z(t).$$

We assume that

(i) the systems are observable in the sense that no two initial states give rise to the same response y for all piecewise continuous inputs, and

(ii) there exist one-to-one maps, say $c(\cdot)$ and $h(\cdot)$, both mapping into a set S such that if each system starts at the identity state and if each system receives the same input, $c(\mathscr{C}X(t)) = h(\mathscr{H}Z(t))$ for all future time. A pair of systems meeting these criteria will be said to be *observable realizations of the same input–output map.* We emphasize that $X(t)$ and $Z(t)$ are square matrices but not necessarily of the same dimension.

Suppose we have two observable realizations of the same input–output map. Let $u(\cdot)$ be a nonzero piecewise constant control defined on $[0, 1]$ which when applied to the X system takes the state $X(0) = I$ into the state $X(1) = I$. Then of course it must do the same for the Z system because they are observable realizations of the same input–output map. Thus we see that if $i_1, i_2, \cdots, i_q$ is a collection of integers with $1 \leqq i_k \leqq m$ and if α_i are any real numbers such that

$$e^{\alpha_1 B_{i_1}} e^{\alpha_2 B_{i_2}} \cdots e^{\alpha_q B_{i_q}} = I,$$

then

$$e^{\alpha_1 G_{i_1}} e^{\alpha_2 G_{i_2}} \cdots e^{\alpha_q G_{i_q}} = I.$$

Let $L_1, L_2, \cdots, L_r$ be a set of commutator expressions in $B_1, B_2, \cdots, B_m$ such that $\{L_i\}$ forms a basis for $\{B_i\}_A$. Let $K_1, K_2, \cdots, K_r$ be in an analogous expression obtained by replacing B_1 by G_1, B_2 by G_2, etc. Let S be an arbitrary commutator expression in $B_1, B_2, \cdots, B_m$ and let T be the analogous commutator expression in $G_1, G_2, \cdots, G_m$. Then in the notation of the proof of Theorem 1, there exist differentiable functions $\alpha_i(\rho)$ such that for $|\rho|$ small,

$$\text{EXP } \alpha_1(\rho)L_1 \text{ EXP } \alpha_2(\rho)L_2 \cdots \text{EXP } \alpha_r(\rho)L_r = \text{EXP } \rho S$$

and

$$\text{EXP } \alpha_1(\rho)K_1 \text{ EXP } \alpha_2(\rho)K_2 \cdots \text{EXP } \alpha_r(\rho)K_r = \text{EXP } \rho T.$$

Since the α_i are differentiable we can write (prime denotes derivative)

$$I + \rho \sum_{i=1}^{m} \alpha_i'(0)L_i + o(\rho^2) = I + \rho S + o(\rho^2)$$

and

$$I + \rho \sum_{i=1}^{m} \alpha_i'(0)K_i + o(\rho^2) = I + \rho T + o(\rho^2).$$

Thus if

$$S = \sum_{i=1}^{r} \gamma_i L_i,$$

then

$$T = \sum_{i=1}^{r} \gamma_i K_i.$$

From this we see that the algebra $\{G_i\}_A$ is generated from $\{B_i\}$ in exactly the same way as the algebra $\{B_i\}_A$ is generated from $\{B_i\}$ and thus that the algebras are isomorphic. We summarize this discussion with a theorem.

THEOREM 10. *Consider the two systems*

$$\dot{X}(t) = \left(\sum_{i=1}^{m} u_i(t) B_i X(t) \right), \qquad y(t) = \mathscr{C} X(t),$$

$$\dot{Z}(t) = \left(\sum_{i=1}^{m} u_i(t) G_i Z(t) \right), \qquad y(t) = \mathscr{H} Z(t),$$

where X and Z are $n \times n$ and $q \times q$ respectively. Suppose that these systems are observable realizations of the same input–output map. Then $\{B_i\}_A$ and $\{G_i\}_A$ are isomorphic as Lie algebras, and moreover if $L_1, L_2, \cdots, L_r$ are commutator expressions in $\{B_i\}$ which form a basis for $\{B_i\}_A$ and if $K_1, K_2, \cdots, K_r$ are the analogous expressions in G_i obtained by replacing B_i by G_i, then $K_1, K_2, \cdots, K_r$ is a basis for $\{G_i\}_A$, and if

$$[L_i, L_j] = \sum_{k=1}^{r} \gamma_{ijk} L_k,$$

then

$$[K_i, K_j] = \sum_{k=1}^{r} \gamma_{ijk} K_k.$$

Of course this does not mean that the reachable sets from I, namely $\{\exp\{B_i\}_A\}_G$ and $\{\exp\{G_i\}_A\}_G$, are isomorphic as groups. For example, the group of unit quaternions and the group of 3×3 orthogonal matrices have isomorphic Lie algebras, yet they are not isomorphic as groups.

7. System theory on coset spaces. In this section we reinterpret our results in a somewhat different way. This interpretation leads to some facts about systems on manifolds which do not admit a group structure. In particular we have in mind the n-sphere $S^n = \{x : x'x = 1, x \in \mathbb{R}^{n+1}\}$ which, as is well known, does not admit a Lie group structure except for the cases $n = 1$ and 3.

Let $M \subset \mathbb{R}^n$ be a manifold. Let $\mathscr{G}$ be a matrix group in $\mathbb{R}^{n \times n}$. We say that $\mathscr{G}$ *acts on* M if for every $x \in M$ and every $G \in \mathscr{G}$, Gx belongs to M. By the *orbit* of $\mathscr{G}$ through x we mean the set of points $\mathscr{G}x = \{y : y = Gx, G \in \mathscr{G}\}$. We say that $\mathscr{G}$ acts *transitively* on M if it acts on M and if for every pair of points x, y in M, there exists G in $\mathscr{G}$ such that $Gx = y$. If $\mathscr{G}$ acts transitively on M, then at any point $x \in M$ there will be a subset $\mathscr{H}_x \subset \mathscr{G}$ such that for each $H \in \mathscr{H}_x$, $Hx = x$. Clearly if $H_1 \in \mathscr{H}_x$ and $H_2 \in \mathscr{H}_x$, then $H_1 H_2 x = H_1 x = x$ and $H^{-1}x = x$ so that $\mathscr{H}_x$ is a subgroup.

We call $\mathscr{H}_x$ the *isotropy group* at x. Notice that if $Gx = y$, then $y = G\mathscr{H}_x x = G\mathscr{H}_x G^{-1} y$, and thus $G\mathscr{H}_x G^{-1}$ is the isotropy group at y—all isotropy groups are conjugate in $\mathscr{G}$. Now suppose M is a manifold for which there actually exists a group $\mathscr{G}$ acting transitively. Pick a point $x \in M$. Define in $\mathscr{G}$ an equivalence relation whereby $G_1 \sim G_2$ if and only if $G_1 = G_2 H_x$ for some $H_x \in \mathscr{H}_x$. There is a one-to-one correspondence between this space of equivalence classes, $\mathscr{G}/\mathscr{H}_x$, and M. In this case we call M a *coset space.*

We study systems in which the state is represented as an n-vector and the evolution is governed by

$$\dot{x}(t) = \left(A + \sum_{i=1}^{v} u_i(t) B_i\right) x(t), \qquad y(t) = \mathscr{C} x(t). \tag{1}$$

By $\mathscr{C}x(t)$ we mean an equivalence class of vectors, x_1 being equivalent to x_2 if and only if $Cx_1 = x_2$ for some C in $\mathscr{C}$.

Let $\mathscr{L}$ be a Lie algebra generated by $\{A, B_i\}$ and let $M \subset \mathbb{R}^n$ be a manifold such that $\{\exp \mathscr{L}\}_G$ acts on it. Then the above equation can be thought of as evolving on the manifold $M \subset \mathbb{R}^n$, for if $x(0) \in M$, then regardless of the control, $x(t) \in M$ for all $t > 0$. If there exists a differentiable manifold $M \subset \mathbb{R}^n$ such that $\{\exp \mathscr{L}\}_G$ acts on M, then we shall say that (1) is *well-posed* on M.

Example 6. Consider the n-sphere, S^n. Let $B_1, B_2, \cdots, B_m$ be $(n+1) \times (n+1)$ skew symmetric matrices. Clearly, the system

$$\dot{x}(t) = \left[\sum_{i=1}^{m} u_i(t) B_i\right] x(t), \qquad y(t) = \mathscr{C} x(t),$$

is well-posed on S^n since $\{\exp \mathscr{L}\}_G$ consists of orthogonal matrices, and orthogonal transformations preserve norm. If we can observe only the first component of x, then we should let $\mathscr{C}$ be the subsets of $\mathscr{SO}(n+1)$ consisting of those matrices which have a 1 in the first column and first row. That is,

$$\mathscr{C} = \begin{bmatrix} 1 & 0 \\ 0 & \mathscr{SO}(n) \end{bmatrix}.$$

With respect to controllability we can say given any two vectors x_1 and x_2 in S^n there exists a piecewise continuous control which steers the system from x_1 to x_2 if and only if $x_2 = Rx_1$ for some R in $\{\exp \mathscr{L}\}_G$, where $\mathscr{L}$ is the Lie algebra generated by $\{A_i\}$. Also, an arbitrary point can be transferred to an arbitrary point if and only if $\{\exp \mathscr{L}\}_G$ acts transitively on S^n.

At the same time we might observe that any x_0 such that $\|x_0\| = 1$ can be transferred to any x_1 such that $\|x_1\| = 1$ if and only if $\{\exp \mathscr{L}\}_G$ acts transitively on S^n. This second point of view is useful because it puts the problem of controllability on S^n in contact with standard results in geometry. In particular a great deal is known about Lie groups which act transivitely on S^n. (See Samelson [17, p. 26].)

As for observability, we note that two initial states x_1 and x_2 in S^n give rise to the same y if and only if for all R in $\{\exp \mathscr{L}\}_G$ there exists $C(R)$ in $\mathscr{C}$ such that $Rx_1 = C(R)Rx_2$, which is to say that $R^{-1}C(R)Rx_2 = x_1$.

We now abstract from this example the essential features and state formally a result which summarizes the development.

THEOREM 11. *Consider the dynamical system* $(x(t) \in \mathbb{R}^n)$

$$\dot{x}(t) = \left(\sum_{i=1}^{q} u_i(t) B_i \right) x(t), \qquad y(t) = \{\exp \mathscr{H}\}_G x(t),$$

which is well-posed on the manifold $M \subset \mathbb{R}^n$. *Let* $\mathscr{L}$ *be the Lie algebra generated by* $\{B_i\}$. *A given state* x_2 *is reachable from* x_1 *if and only if* $x_2 = Nx_1$ *for some* N *in* $\{\exp \mathscr{L}\}_G$. *Let* $\mathscr{P} = \{P : RPR^{-1} \in \{\exp \mathscr{H}\}_G$ *for all* $R \in \exp \mathscr{L}\}$. *Two states* x_1 *and* x_2 *are indistinguishable if and only if* $x_2 = Px_1$ *for some* P *in* $\mathscr{P}$. *In particular, two states* x_1 *and* x_2 *are indistinguishable if* $x_2 = Px_1$ *for* P *in* $\{\exp \mathscr{K}\}_G$ *with* $\mathscr{K}$ *being any Lie algebra such that* $\{\mathrm{ad}_{\mathscr{L}} \mathscr{K}\}_A \subset \mathscr{K}$.

Example 7. Consider the submanifold M of $\mathbb{R}^{n+1}$ consisting of those points whose last coordinate is 1. The evolution equation in M,

$$\frac{d}{dt}\begin{bmatrix} x(t) \\ 1 \end{bmatrix} = \begin{bmatrix} A & Bu(t) \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x \\ 1 \end{bmatrix}, \qquad y(t) = \left(\exp \begin{bmatrix} 0 & \ker C \\ 0 & 0 \end{bmatrix} \right)\begin{bmatrix} x(t) \\ 1 \end{bmatrix}$$

corresponds to the more familiar $\dot{x}(t) = Ax(t) + Bu(t), y(t) = Cx(t)$. Using Theorem 7 we see that for the associated group equation

$$\dot{X}(t) = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} + \sum_{i=1}^{m} u_i(t) \begin{bmatrix} 0 & b_i \\ 0 & 0 \end{bmatrix} X(t)$$

the reachable set at time t consists of those matrices which can be written as

$$\mathscr{R} = \left\{ X : X = \begin{bmatrix} e^{At} & x \\ 0 & 1 \end{bmatrix};\ x \in \mathrm{range}\,(B, AB, \cdots, A^{n-1}B) \right\}.$$

Thus if $B, AB, \cdots, A^{n-1}B$ spans $\mathbb{R}^n$, then the reachable group acts transitively on M and we have controllability.

As for observability, we note that

$$\exp \begin{pmatrix} 0 & \ker C \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & \ker C \\ 0 & 1 \end{pmatrix}.$$

The subalgebras of $\mathscr{H}$ which are closed under commutation with $\mathscr{L}$ correspond to the linear subspaces of $\ker C$ which are invariant under A.

Acknowledgment. Among the many people who made useful suggestions on earlier expositions I want to mention in particular D. Elliott, V. Jurdjevic, H. Sussmann, Jacques Willems and Jan Willems. I also want to thank H. Rosenbrock who some years ago acquainted me with work relating to differential equations and Lie algebras.

REFERENCES

[1] W. L. CHOW, *Über System von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann., 117 (1939), pp. 98–105.

[2] R. HERMANN, *On the accessibility problem in control theory*, Proc. International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.

[3] J. KUCERA, *Solution in large of control problem*: $\dot{x} = (A(1 - u) + Bu)x$, Czech. Math. J., 16 (1966), no. 91, pp. 600–623.

[4] J. KUCERA, *Solution in large of control problem: $\dot{x} = (Au + Bv)x$*, Czech. Math. J., 17 (1967), no. 92, pp. 91–96.
[5] C. LOBRY, *Contrôlabilité des systèmes non linéaries*, this Journal, 8 (1970), pp. 573–605.
[6] G. W. HAYNES AND H. HERMES, *Nonlinear controllability via Lie theory*, this Journal, 8 (1970), pp. 450–460.
[7] W. MAGNUS, *On the exponential solution of differential equations for a linear operator*, Comm. Pure Appl. Math., 7 (1954), pp. 649–673.
[8] K. T. CHEN, *Decomposition of differential equations*, Math. Ann., 146 (1962), pp. 263–278.
[9] J. WEI AND E. NORMAN, *On global representations of the solutions of linear differential equations as a product of exponentials*, Proc. Amer. Math. Soc., 15 (1964), pp. 327–334.
[10] ———, *Lie algebraic solution of linear differential equations*, J. Mathematical Phys., 4 (1963), pp. 575–581.
[11] I. M. SINGER AND J. A. THORPE, *Lecture Notes on Elementary Topology and Geometry*, Scott, Foresman and Co., Glenview, Ill., 1967.
[12] S. STERNBERG, *Lectures on Differential Geometry*, Prentice-Hall, Englewood Cliffs, N.J., 1964.
[13] L. AUSLANDER AND R. MACKENZIE, *Introduction to Differentiable Manifolds*, McGraw-Hill, New York, 1963.
[14] J. J. ROTMANN, *The Theory of Groups*, Allyn and Bacon, Boston, 1965.
[15] J. F. P. MARTIN, *Some results on matrices which commute with their derivatives*, SIAM J. Appl. Math., 15 (1967), pp. 1171–1183.
[16] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
[17] H. SAMELSON, *Topology of Lie groups*, Bull. Amer. Math. Soc., 58 (1952), pp. 2–37.
[18] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, to appear.
[19] V. JURDJEVIC AND H. J. SUSSMANN, *Control systems on Lie groups*, J. Differential Equations, to appear.
[20] D. L. ELLIOTT, *A consequence of controllability*, J. Differential Equations, 10 (1971), pp. 364–370.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

STOCHASTIC CONTROL: A FUNCTION SPACE APPROACH*

A. V. BALAKRISHNAN†

Abstract. First results in a new systematic approach using function space techniques on stochastic control problems based on imperfect observations, providing in particular an extension of the separation principle, are given.

1. Introduction. Stochastic control problems even for linear systems have for the most part so far been studied by using dynamic programming, the work of Wonham [1] and Kushner [2] being undoubtedly the most notable. The recent survey article by Flemming [3] presents an excellent review. In this paper we present the first results in a new approach using function space techniques much in the same manner as in dealing with nonstochastic control problems (e.g., [4], [5]). Apart from the obvious benefits of unification of methods, the approach has other advantages. It sheds more light on the structure of stochastic controls based on imperfect observation and the so-called "separation" principle separating filter and control. It also facilitates the extension to partial differential equations (although not treated here).

Essential use is made of the "innovation" process whose importance has been brought out in the recent work of Kailath [6] in filtering theory. Our own approach to Kalman filtering is in [7]. We only consider linear systems and only one class of problems with quadratic cost, owing to space limitation. Other cases, including differential games, are treated in [9].

The bulk of the paper is devoted to setting up the appropriate Hilbert spaces of controls based on observation. The technique for determining optimal controls is described in § 3.

2. Hilbert spaces of admissible controls. The system description in the usual notation will be taken as

$$x(t;\omega) = \int_0^t A(s)x(s;\omega)\,ds + \int_0^t B(s)u(s;\omega)\,ds + \int_0^t F(s)\,dW(s;\omega), \tag{2.1}$$

$$Y(t;\omega) = \int_0^t C(s)x(s;\omega)\,ds + \int_0^t G(s)\,dW(s;\omega), \qquad 0 \leqq t \leqq 1, \tag{2.2}$$

where

$x(t;\omega)$ is an $n \times 1$ "state" function,
$u(t;\omega)$ is an $m \times 1$ "control" function,
$W(t;\omega)$ is a $q \times 1$ Wiener process,

* Received by the editors September 24, 1971, and in final revised form December 30, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Systems Science Department, University of California, Los Angeles, California 90024. This work was supported in part by the Office of Scientific Research, U.S. Air Force, Applied Mathematics Division, under Grant 68-1408.

ω = (sample) points in the B-space $C[0, 1]^q$ (space of $q \times 1$ matrix functions continuous on $[0, 1]$ with sup norm), which space we shall henceforth denote by C,

$\mathscr{B}$ = the sigma algebra of Borel sets in C,

$\mathscr{B}(t)$ = the sigma algebra generated by $W(s;\omega)$, $0 \leqq s \leqq t$,

and the coefficients

$$\underset{(n \times n)}{A(s)}, \quad \underset{(n \times m)}{B(s)}, \quad \underset{(n \times q)}{F(s)}, \quad \underset{(q \times n)}{C(s)}, \quad \underset{(p \times q)}{G(s)}$$

are all taken (for simplicity) to be continuous on $[0, 1]$.

Under the blanket assumption imposed in all that follows that the controls $u(t;\omega)$ are jointly measurable on $t \times \omega$ (Lebesgue measure on t) and further that

$$\int_0^1 E\|u(t;\omega)\|^2 \, dt < \infty, \qquad u(t;\omega) \text{ measurable } \mathscr{B}(t) \text{ a.e.}, \tag{2.3}$$

it is well known that (2.1) has a unique solution which is continuous in t for almost all ω.

The class of control problems we shall consider is then given as follows: let $\mathscr{B}_Y(t)$ denote the sigma algebra generated by $Y(s;\omega)$, $0 \leqq s \leqq t$; it is required to minimize the cost functional

$$\int_0^1 E[g(t; x(t;\omega); u(t;\omega))] \, dt \tag{2.4}$$

(where the real-valued function $g(\cdot;\cdot;\cdot)$ is assumed to be continuous in all the variables) in the class of controls satisfying in addition the crucial restriction that $u(t;\omega)$ be measurable $\mathscr{B}_Y(t)$, a.e. in t, $0 \leqq t \leqq 1$. The class of such controls is obviously a linear class; let us denote it by $\mathscr{H}_Y$. Let us denote the Hilbert space of controls satisfying (2.3) by $\mathscr{H}$. Then $\mathscr{H}_Y$ is a (not necessarily closed) subspace of $\mathscr{H}$, and (2.4) defines a functional on $\mathscr{H}_Y$. The circular dependence of $u(t;\omega)$ on $\mathscr{B}_Y(t)$ makes it hard to characterize the space $\mathscr{H}_Y$ in a way useful for determining optimal controls, and therein is our first problem. For this purpose we proceed in the following way, invoking results from (Kalman) filtering theory. Thus let

$$x_u(t;\omega) = \int_0^t A(s)x_u(s;\omega)\, ds + \int_0^t B(s)u(s;\omega)\, ds,$$

$$Y_u(t;\omega) = \int_0^t C(s)x_u(s;\omega)\, ds,$$

$$\tilde{x}(t;\omega) = x(t;\omega) - x_u(t;\omega),$$

$$\tilde{Y}(t;\omega) = Y(t;\omega) - Y_u(t;\omega).$$

We shall assume throughout that

$$G(s)G(s)^* > 0, \qquad 0 \leqq s \leqq 1. \tag{2.5}$$

Let $\mathscr{B}_{\tilde{Y}}(t)$ denote the sigma algebra generated by $\tilde{Y}(s;\omega)$, $0 \leqq s \leqq t$, and let

$$\hat{\tilde{x}}(t;\omega) = E(\tilde{x}(t;\omega)|\mathscr{B}_Y(t)), \qquad \hat{x}(t;\omega) = \hat{\tilde{x}}(t;\omega) + x_u(t;\omega).$$

Then $\tilde{x}(t;\omega)$ satisfies (2.1) with $u(t;\omega)$ set identically to be zero, and $\tilde{Y}(t;\omega)$ is given by (2.2) with $\tilde{x}(t;\omega)$ in place of $x(t;\omega)$. Hence we have (see [1], [6] and [7] for a more recent treatment using Martingales)

$$\hat{\tilde{x}}(t;\omega) = \int_0^t A(s)\hat{\tilde{x}}(s;\omega)\,ds + \int_0^t K(s)\,dZ(s;\omega), \tag{2.6}$$

where

$$K(s) = (P(s)C(s)^* + F(s)G(s)^*)(G(s)G(s)^*)^{-1},$$

$$\dot{P}(s) = A(s)P(s) + P(s)A(s)^* + F(s)F(s)^* - (P(s)C(s)^* + F(s)G(s)^*)(G(s)G(s)^*)^{-1}(C(s)P(s) + G(s)F(s)^*), \qquad P(0) = 0,$$

and

$$Z(t;\omega) = \tilde{Y}(t;\omega) - \int_0^t C(s)\hat{\tilde{x}}(s;\omega)\,ds \tag{2.7}$$

and $Z(s;\omega)$ is a Gaussian Martingale with covariance $G(s)G(s)^*$:

$$E(Z(t;\omega)|\mathscr{B}_Z(s)) = Z(s;\omega), \qquad s \leqq t,$$

$$E[(Z(t_2;\omega) - Z(t_1;\omega))(Z(t_2;\omega) - Z(t_1;\omega))^*|\mathscr{B}_Z(t_1)] = \int_{t_1}^{t_2} G(s)G(s)^*\,ds,$$

where $\mathscr{B}_Z(t)$ denotes the sigma algebra generated by $Z(s;\omega)$, $0 \leqq s \leqq t$. Moreover,

$$\mathscr{B}_Z(t) = \mathscr{B}_{\tilde{Y}}(t). \tag{2.8}$$

Note now the obvious but important relations that

$$Z(t;\omega) = Y(t;\omega) - \int_0^t C(s)\hat{x}(s;\omega)\,ds, \tag{2.9}$$

$$x(t;\omega) = \hat{x}(t;\omega) + e(t;\omega), \tag{2.10}$$

$e(t;\omega) \equiv \tilde{x}(t;\omega) - \hat{\tilde{x}}(t;\omega)$ and is independent of the sigma algebra $\mathscr{B}_Z(t)$, and of course that $e(t;\omega)$ is Gaussian with covariance $P(t)$. Moreover,

$$\hat{x}(t;\omega) = \int_0^t A(s)\hat{x}(s;\omega)\,ds + \int_0^t K(s)\,dZ(s;\omega) + \int_0^t B(s)u(s;\omega)\,ds. \tag{2.6a}$$

Note that $Y(t;\omega)$ is an R_2-process in the terminology of Nelson [8], and we can define the Itô integral:

$$\int_0^1 [f(s), dY(s;\omega)] = \int_0^1 [f(s), C(s)\hat{x}(s;\omega)]\,ds + \int_0^1 [f(s), dZ(s;\omega)]. \tag{2.11}$$

Our first observation now is that any control $u(t;\omega)$ in $\mathscr{H}$ such that $u(t;\omega)$ is measurable $\mathscr{B}_Z(t)$ for almost every t in $[0, 1]$ is actually in $\mathscr{H}_Y$. We shall state this as the following theorem.

THEOREM 2.1. *Any control function $u(t;\omega)$ in $\mathscr{H}$ such that $u(t;\omega)$ is measurable $\mathscr{B}_Z(t)$ a.e., in t is actually in $\mathscr{H}_Y$.*

Proof.[1] We shall show that

$$\mathscr{B}_Z(t) = \mathscr{B}_Y(t), \qquad 0 \leqq t \leqq 1. \tag{2.12}$$

First of all, let us invoke the representation theorem due to Cameron–Martin–Itô [10]. Without loss of generality, we may take $u(t;\omega)$ centered and $Z(t;\omega)$ to be a Wiener process (covariance equals the identity matrix). For each t (omitting a set of Lebesgue measure zero), there exists a sequence of orthogonal variables $G_n(t;\omega)$ such that

$$\lim_m E\left(\left\| u(t;\omega) - \sum_1^m G_n(t;\omega)\right\|^2\right) = 0,$$

$$G_n(t;\omega) = I_n(K_n;Z;\omega;t),$$

$$I_n(K_n;Z;\omega;t) = \int_0^t \int_0^t K_n(t;s_1,\cdots,s_n;dZ(s_1;\omega),\cdots,dZ(s_n;\omega)),$$

where $K_n(t;s_1,\cdots,s_n;x_1,\cdots,x_n)$ is a symmetric n-linear form in the x_i for each t, $s_1,\cdots,s_n$, and jointly Lebesgue measurable $\{s_i\}$; $K_n(t;s,s,\cdots,s;\cdot)$ is zero, and symmetric in the s_i and

$$E(\|I_n(K_n;Z;\omega;t)\|^2) = n!\int_0^t\cdots\int_0^t \|K_n(t;s_1,\cdots,s_n;\cdots)\|^2\, ds_1\cdots ds_n$$
$$= n!\,\|K_n(t;\cdot)\|^2.$$

Further, for any symmetric n-linear kernel $f(s_1,\cdots,s_n;x_1,\cdots,x_n)$ with $f(s,\cdots,s;x_1,\cdots,x_n)$ zero and symmetric in the s_i and

$$\int_0^1\cdots\int_0^1 \|f(s_1,\cdots,s_n;\cdots)\|^2\, ds_1\cdots ds_n < \infty,$$

$$E([u(t;\omega), I_n(f;Z;\omega;1)])$$
$$= \int_0^t\cdots\int_0^t [K_n(t;s_1,\cdots,s_n;\cdots), f(s_1,\cdots,s_n;\cdots)]\, ds_1\cdots ds_n. \tag{2.13}$$

From the given joint measurability of t and ω of $u(t;\omega)$, it follows by using (2.13) that $K_n(t,s_i;\cdot)$ is Lebesgue measurable in t. Further, we have

$$\int_0^1 E(\|u(t;\omega)\|^2)\, dt = \int_0^1 \sum_1^\infty n!\|K_n(t;\cdot)\|^2\, dt < \infty,$$

implying that $K_n(t;s_i;\cdot)$ is (or can be taken to be) jointly Lebesgue measurable in t, s_i, and

$$u(t;\omega) = \sum_1^\infty I_n(K_n;Z;\omega;t), \tag{2.14}$$

where the covergence is in the norm of $\mathscr{H}_Z$.

[1] I am indebted to Yu. Rozanov for pointing out a lacuna in an earlier version of the proof.

Let us note that the kernels in (2.14) can be obtained directly in the following way. Consider the (closed) subspace in $\mathscr{H}_Z$ of elements of the form

$$I_n(f; Z; \omega; t),$$

where $f(t; s_1, \cdots, s_n; \cdot)$ is a symmetric n-linear form jointly measurable in the variables t, s_i; $f(t; s, \cdots, s; \cdot)$ is zero and

$$\int_0^1 dt \int_0^t \cdots \int_0^t \| f(t; s_1, \cdots, s_n; \cdot)\|^2 \, ds_1 \cdots ds_n < \infty.$$

Then

$$\int_0^1 E([u(t, \omega), I_n(f; Z; \omega; t)]) \, dt$$

defines a continuous linear functional, and hence by the Riesz theorem there exists $K_n(t; s_i; \cdot)$ such that the above integral is equal to

$$n! \int_0^1 dt \int_0^t \cdots \int_0^t [K_n(t; s_1, \cdots, s_n; \cdot), f(t; s_1, \cdots, s_n)] \, ds_1 \cdots ds_n$$
$$= \int_0^1 E([I_n(K_n; Z; \omega; t), I_n(f; Z; \omega; t)]) \, dt.$$

Moreover, as in Itô [10], we can, by iterated integration, obtain the representations

$$I_n(K_n; Z; \omega; t) = \int_0^t k_n(t; s; \omega) \, dZ(s; \omega),$$

where $k_n(t; s; \omega)$ is measurable $\mathscr{B}_Z(s)$ for $s \leqq t$, and jointly measurable in all the variables. Finally using (2.14), we have the representation

$$u(t; \omega) = \int_0^t K(t; s; \omega) \, dZ(s; \omega), \tag{2.15}$$

where $K(t; s; \omega)$ is measurable $\mathscr{B}_Z(s)$, and jointly measurable in t, s and ω, and

$$\int_0^1 \int_0^t E(\| K(t; s; \omega)\|^2) \, ds \, dt < \infty.$$

Finally we have, using (2.6a), the representation

$$C(t)\hat{x}(t; \omega) = \sum_1^\infty I_n(\tilde{K}_n; Z; \omega; t), \tag{2.16}$$

where

$$\sum_1^\infty \int_0^1 n! \| \tilde{K}_n(t; \cdot)\|^2 \, dt < \infty.$$

We shall show that $\mathscr{B}_Y(t) = \mathscr{B}_Z(t)$ by showing that for any $f(\cdot)$ in $L_2(0, t)^p$,

$$\int_0^t [f(s), dZ(s; \omega)] \quad \text{is measurable } \mathscr{B}_Y(t).$$

We shall find an explicit representation; and to see why it works, we shall first

indicate briefly the general line of reasoning. Thus let H_n denote the L_2-space of symmetric n-linear forms $h_n(s; x_1, \cdots, x_n)$, $x_i \in E_p$, with range in E_1, $s = (s_1, \cdots, s_n)$, $0 \leqq s_i < t$, and

$$\int_0^t \cdots \int_0^t \|h_n(s; \cdots)\|^2 \, d|s| = \|h_n\|^2.$$

Let H_∞ denote the infinite product space with norm defined by

$$\|h\|^2 = \sum_1^\infty n! \|h_n\|^2.$$

For any function $z(\cdot)$ in $L_2(0, t)^p$, define (with $\tilde{K}_n(\cdot)$ as in (2.16))

$$y = P(z); \; y(s) = \sum_1^\infty \int_0^s \int_0^s \tilde{K}_n(s; s_1, \cdots, s_n; z(s_1), \cdots, z(s_n)) \, ds_1 \cdots ds_n.$$

Note that

$$\int_0^t \|y(s)\|^2 \, ds \leqq \left(\sum n! \int_0^t \|\tilde{K}_n(s; \cdot)\|^2 \, ds \right) \left(\sum \|z\|^{2n}/n! \right),$$

so that $P(\cdot)$ is a nonlinear (analytic) Volterra operator mapping $L_2(0, t)$ into itself. Next for each $z(\cdot)$ in $L_2(0, t)$ and each kernel h, we can define the (continuous, nonlinear!) functional

$$[h, z] = \sum_1^\infty \int_0^t \cdots \int_0^t h_n(s_1, \cdots, s_n; z(s_1), \cdots, z(s_n)) \, ds_1 \cdots ds_n,$$

where the $h_n(s_1, \cdots, s_n; \cdot)$ may be taken to be symmetric in the variables s_i as well. Then the main point is that we can verify, as we do in greater detail below,

$$[h, z + P(z)] = [h + L(h), z],$$

where L is a linear Volterra operator mapping H_∞ into itself. Hence

$$[h, z] = [(I + L)^{-1} h, z + P(z)],$$

where in our application we need to consider the special case

$$h = \{g_k\}, \quad g_k = 0, \quad k > 1.$$

We have also to deal with random processes, instead of functions, but the result is purely function theoretic in nature.

Now the main calculation is:

$$\begin{aligned} &\int_0^t \cdots \int_0^t h_n(s_1, \cdots, s_n; dY(s_1; \omega), \cdots, dY(s_n; \omega)) \\ &\quad = I_n(h_n; Z; \omega) + \sum_{j=n}^\infty I_j(L_{j,n}(h_n); Z; \omega), \end{aligned} \tag{2.17}$$

where

$$I_j(h_j; Z; \omega) = \int_0^t \cdots \int_0^t h_j(s_1, \cdots, s_j; dZ(s_1; \omega), \cdots, dZ(s_j; \omega))$$

and $L_{j,n}$ is a linear Volterra operator mapping H_n into H_j. In order to conserve

notation, we shall show this for $n = 2$. Thus

$$\int_0^t \cdots \int_0^t h_2(s_1, s_2; dY(s_1; \omega), dY(s_2; \omega))$$

$$= I_2(h_2; Z; \omega) + 2\int_0^t \cdots \int_0^t h(s_1, s_2; C(s_1)\hat{x}(s_1; \omega); dZ(s_2; \omega))\, ds_1 \tag{2.18}$$

$$+ \int_0^t \cdots \int_0^t h(s_1, s_2; C(s_1)\hat{x}(s_1; \omega); C(s_2)\hat{x}(s_2; \omega))\, ds_1\, ds_2.$$

Now the second term is expanded into an infinite series, a typical term of which is:

$$\int_0^t \cdots \int_0^t h\left(s_1, s_2; \int_0^s \cdots \int_0^s \tilde{K}_n(s_1; \sigma_1, \cdots, \sigma_n; dZ(\sigma_1; \omega), \cdots, dZ(\sigma_n; \omega)); dZ(s_2)\right) ds_1$$

which we can rewrite as

$$\int_0^t \cdots \int_0^t k_{n+1}(\sigma_1, \cdots, \sigma_n, s_2; dZ(\sigma_1), \cdots, dZ(\sigma_n); dZ(s_2)),$$

where

$$k_{n+1}(\sigma_1, \cdots, \sigma_n, s_2; \cdot) = \int_{\max \sigma_i}^t h(s_1, s_2; \tilde{K}_n(s_1; \sigma_1, \cdots, \sigma_n; \cdots), \cdot)\, ds_1.$$

Thus defined, we have clearly a (linear) Volterra operator mapping H_2 into H_{n+1}, and

$$\|k_{n+1}\|^2 \leqq \|h\|^2 \int_0^t \|\tilde{K}_n(s; \cdots; \cdots)\|^2\, ds$$

and similarly then for the other terms in (2.18). Moreover, combined with (2.16) we see further that

$$\sum_2^\infty j!\|L_{j,2}(h_2)\|^2/\|h_2\|^2 < \infty,$$

so that in particular the series in (2.17) converges in the mean square. Hence finally for h in H_∞, we obtain

$$\sum_1^\infty \int_0^t \cdots \int_0^t h_n(s_1, \cdots, s_n; dY(s_1; \omega), \cdots, dY(s_n; \omega))$$

$$= \sum_{n=1}^\infty \left[I_n(h_n + L_{n,n}(h_n); Z; \omega) + \sum_{j=n+1}^\infty I_j(L_{j,n}(h_n); Z; \omega) \right] \cdots.$$

We shall now "invert" this equation for the case where the right side is equal to

$$\int_0^t [f(s), dZ(s; \omega)] = \int_0^t f_1(s; dZ(s; \omega)).$$

We readily see that

$$(I + L_{1,1})h_1 = f_1,$$

$$(I + L_{2,2})h_2 + L_{2,1}(h_1) = 0,$$

$$(I + L_{k,k})h_k + \sum_{j=k-1}^1 L_{k,j}(h_j) = 0;$$

and because the $L_{k,k}$ are Volterra, we see that "triangular matrix" equations are readily solved for the $h_k(\cdot)$, yielding

$$\int_0^t f_1(s; dZ(s;\omega)) = \sum_{k=1}^{\infty} \int_0^t \cdots \int_0^t h_k(s_1, \cdots, s_k; dY(s_1;\omega), \cdots, dY(s_k;\omega)).$$

Since from (2.17) we have

$$E\left(\int_0^t \cdots \int_0^t h_n(s_1, \cdots, s_n; dY(s_1;\omega), \cdots, dY(s_n;\omega))\right)^2$$

$$= n!\|(I + L_{n,n})(h_n)\|^2 + \sum_{n+1}^{\infty} j!\|L_{j,n}(h_n)\|^2,$$

the necessary mean square convergence of the series is readily established. This proves (2.12).

Next, the class of control functions $u(t;\omega)$ in $\mathscr{H}$ such that $u(t;\omega)$ is measurable $\mathscr{B}_Z(t)$ a.e. in t is clearly closed. Let us denote this Hilbert space by $\mathscr{H}_Z$. Then of course

$$\mathscr{H}_Z \subset \mathscr{H}_Y,$$

but whether equality holds would appear to be a moot question. However we can assert the following theorem.

THEOREM 2.2. *With the kernels $K_n(t;\cdot;\cdot)$ of the kind defined in Theorem* 2.1, *suppose $u(t;\omega)$ has the representation*

$$u(t;\omega) = \sum_1^{\infty} a_n(t) I_n(K_n; Y; \omega; t), \tag{2.19}$$

where

$$\sum_1^{\infty} \int_0^1 n!|a_n(t)|^2 \|K_n(t;\cdot)\|^2 \, dt < \infty.$$

Then $u(t;\omega)$ belongs to $\mathscr{H}_Z$.

Proof. Let us observe first that by iterated integration, following Itô, we have again that

$$u(t;\omega) = \int_0^t K(t;s;\omega)\, dY(s;\omega), \tag{2.20}$$

$$\int_0^1 \int_0^t E(\|K(t;s;\omega)\|^2)\, ds\, dt < \infty.$$

We define

$$\tilde{Z}(t;\omega) = Y(t;\omega) - \int_0^t C(s) x_u(s;\omega)\, ds$$

and use the representation

$$C(t)x_u(t;\omega) = \sum_1^{\infty} \int_0^t a_n(t;s) I_n(K_n; Y; \omega; s)\, ds,$$

which we rewrite as before as:

$$C(t)x_u(t;\omega) = \sum_1^\infty I_n(\tilde{K}_n; Y; \omega; t),$$

where

$$\sum_1^\infty \int_0^1 n!\|\tilde{K}_n(t;\cdot)\|^2\, dt < \infty.$$

Then we can calculate as in Theorem 2.1 that

$$\int_0^1 [f(s), dY(s;\omega)] = \sum_{n=1}^\infty \int_0^t \cdots \int_0^t h_n(s_1, \cdots, s_n; dZ(s_1;\omega), \cdots, dZ(s_n;\omega)).$$

The roles of $Y(\cdot)$ and $Z(\cdot)$ are reversed, and the analogous general relation is

$$[h, y - P(y)] = [h - L(h), y].$$

Hence

$$\mathscr{B}_Y(t) = \mathscr{B}_{\tilde{Z}}(t).$$

Next, from the relation

$$\tilde{Z}(t;\omega) = Z(t;\omega) + \int_0^t \phi(t)\phi(s)^{-1}K(s)\, dZ(s;\omega), \tag{2.21}$$

where

$$\dot{\phi}(t) = A(t)\phi(t),$$

it readily follows that

$$\mathscr{B}_{\tilde{Z}}(t) = \mathscr{B}_Z(t),$$

and hence the theorem follows.

Another important property of controls in $\mathscr{H}_Z$ is that

$$\hat{x}(t;\omega) = E(x(t;\omega)|\mathscr{B}_Y(t)), \tag{2.22}$$

and this follows readily from the fact that $u(t;\omega)$ is measurable $\mathscr{B}_Y(t)$ and $\mathscr{B}_Z(t)$.

If we need to deal only with linear transformations on the observation we can state the following theorem.

THEOREM 2.3. *The class $\mathscr{H}_\mathscr{L}$ of controls $u(t;\omega)$ such that*

$$u(t;\omega) = \int_0^t K(t;s)\, dY(s;\omega), \tag{2.23}$$

where

$$\int_0^1 \int_0^t \|K(t;s)\|^2\, ds\, dt < \infty,$$

is closed in $\mathscr{H}_Z$ (and in $\mathscr{H}_Y$). Moreover, we have the representation

$$u(t;\omega) = \int_0^t M(t;s)\, dZ(s;\omega), \tag{2.24}$$

where

$$\int_0^1 \int_0^s \|M(t;s)\|^2 \, ds \, dt < \infty .$$

Also, every control of the form (2.24) *is in* $\mathscr{H}_{\mathscr{L}}$ *as well.*

Proof. From (2.23), we readily calculate, with $\tilde{Z}(t;\omega)$ as in Theorem 2.2, that

$$\tilde{Z}(t;\omega) = Y(t;\omega) + \int_0^t \int_0^s L(s;\sigma) \, dY(\sigma;\omega) \, ds,$$

where

$$\int_0^1 \int_0^s \|L(s;\sigma)\|^2 \, d\sigma \, ds < \infty .$$

Let $h(t)$ denote a $p \times p$ matrix function square integrable on $[0, 1]$. Then we have that

$$\int_0^1 h(t) \, dZ(t;\omega) = \int_0^1 g(t) \, dY(t;\omega) + \int_0^1 h(t) \, dY(t;\omega), \tag{2.25}$$

where

$$g = Qh; \qquad g(t) = \int_t^1 h(s) L(s;t) \, ds. \tag{2.26}$$

But Q is a Hilbert-Schmidt Volterra operator, and hence we have

$$(I + Q)^{-1} = I + R,$$

where R has the form

$$Rf = g, \quad g(t) = \int_t^1 f(s) r(s;t) \, ds, \quad \int_0^1 \int_0^s \|r(s;t)\|^2 \, dt \, ds < \infty .$$

Hence, for any $p \times p$ square integrable function $h(\cdot)$, we have

$$\int_0^1 h(t) \, dY(t;\omega) = \int_0^1 k(t) \, dZ(t;\omega), \qquad k = (I + R)h.$$

Expanding the right side we have

$$\int_0^1 k(t) \, dZ(t;\omega) = \int_0^1 h(t) \, dZ(t;\omega) + \int_0^1 h(t) \int_0^t r(t;s) \, dZ(s;\omega) \, ds \, dt$$

from which it follows that

$$Y(t;\omega) = Z(t;\omega) + \int_0^t ds \int_0^s r(s;\sigma) \, dZ(\sigma;\omega).$$

Finally substituting for $\tilde{Z}(t;\omega)$ in terms of $Z(t;\omega)$ from (2.21), we have that

$$Y(t;\omega) = Z(t;\omega) + \int_0^t ds \int_0^s j(s;\sigma) \, dZ(\sigma;\omega), \qquad \int_0^1 \int_0^t \|j(t;s)\|^2 \, ds \, dt < \infty .$$

Substituting into (2.23), we obtain (2.24). Since from (2.24) we also have

$$\int_0^1 E(\|u(t;\omega)\|^2)\,dt = \int_0^1 \int_0^t \|M(t;s)\|^2\,ds\,dt,$$

we readily see that $\mathscr{H}_{\mathscr{L}}$ is closed. Moreover, the same kind of argument enables us to go from (2.24) to (2.23), concluding the proof of the theorem.

Let us now look at the cost functional. For $u(t;\omega)$ in $\mathscr{B}_Z$, we can obtain a "separation principle," namely,

$$\begin{aligned} E(g(t;x(t;\omega);u(t;\omega))) &= E(g(t;\hat{x}(t;\omega)+e(t;\omega);u(t;\omega))) \\ &= E(E(g(t;\hat{x}(t;\omega)+e(t;\omega);u(t;\omega))/\mathscr{B}_Z(t))). \end{aligned}$$

But $e(t;\omega)$ is independent of the sigma algebra $\mathscr{B}_Y(t) = \mathscr{B}_Z(t)$, while $\hat{x}(t;\omega)$ and $u(t;\omega)$ are both measurable $\mathscr{B}_Z(t)$, and hence,

$$E(g(t;\hat{x}(t;\omega)+e(t;\omega);u(t;\omega)/\mathscr{B}_Z(t))) = h(t;\hat{x}(t;\omega);u(t;\omega))$$

so that we may minimize equivalently instead

$$\int_0^1 E(h(t;\hat{x}(t;\omega);u(t;\omega)))\,dt,$$

where $\hat{x}(t;\omega)$ satisfies,

$$\text{(2.6a)} \qquad \hat{x}(t;\omega) = \int_0^t A(s)\hat{x}(s;\omega)\,ds + \int_0^t B(s)u(s;\omega)\,ds + \int_0^t K(s)\,dZ(s;\omega),$$

where

$$Z(s;\omega) = Y(s;\omega) - \int_0^t C(s)\hat{x}(s;\omega)\,ds.$$

Of course the most important cost functional from the practical point of view is the quadratic functional of the form

$$E[Q(t)\hat{x}(t;\omega), x(t;\omega)] + \lambda E[u(t;\omega), u(t;\omega)], \qquad \lambda > 0, \quad Q(t) \geqq 0.$$

For this case the above reduction becomes

$$E[Q(t)\hat{x}(t;\omega), \hat{x}(t;\omega)] + \lambda E[u(t;\omega), u(t;\omega)] + \operatorname{tr} Q(t)P(t),$$

and the problem is thus reduced to that of minimizing

$$\text{(2.27)} \qquad \int_0^1 E[Q(t)\hat{x}(t;\omega), \hat{x}(t;\omega)]\,dt + \lambda \int_0^1 [u(t;\omega), u(t;\omega)]\,dt$$

with $\hat{x}(t;\omega)$ given by (2.6a) and (2.9).

3. Determination of optimal controls. We shall next see how to determine optimal controls by using function space techniques as in deterministic problems. We shall only consider controls in $\mathscr{H}_Z$. Let

$$\text{(3.1)} \qquad v(t;\omega) = \int_0^t \phi(t)\phi(s)^{-1}K(s)\,dZ(s;\omega).$$

Define the linear bounded transformation L mapping $\mathscr{H}_Z$ into the Hilbert space of $n \times n$ functions $x(t;\omega)$ such that

$$\int_0^1 E(\|x(t;\omega)\|^2)\,dt < \infty$$

by

$$Lu = x, \qquad x(t;\omega) = \int_0^t \phi(t)\phi(s)^{-1}B(s)u(s;\omega)\,ds.$$

Then denoting the function $\hat{x}(t;\omega)$ by $\hat{x}$ for short, and $u(t;\omega)$ by u, and using v to denote the function $v(t;\omega)$, we have

$$\hat{x} = Lu + v.$$

Using $[\cdot,\cdot]$ to denote inner product in appropriate Hilbert spaces, and using Q to denote the operator corresponding to multiplication by $Q(t)$ which, for simplicity, we assume to be continuous in $[0, 1]$, we can express the quadratic cost functional (2.27) as a continuous quadratic form over $\mathscr{H}_Z$ by

$$[Q(Lu + v), Lu + v] + \lambda[u, u]. \tag{3.2}$$

As is immediate therefrom, the minimum is attained at a unique point u_0 in $\mathscr{H}_Z$ given by

$$u_0 = (-1/\lambda)L^*Q(Lu_0 + v),$$

where L^* is the adjoint of L. We recognize $Lu_0 + v$ to be the state estimate function $\hat{x}(t;\omega)$ corresponding to the choice $u_0(t;\omega)$. The main problem is thus that of determining the function $L^*Q\hat{x}$.

For this let us first determine L^*x for any x. We must have

$$[L^*x, u] = [x, Lu] = \int_0^1 E\left[x(t;\omega), \int_0^t \phi(t)\phi(s)^{-1}B(s)u(s;\omega)\,ds\right]dt$$

$$= \int_0^1 E\left[\int_s^1 B(s)^*\phi(s)^{*-1}\phi(t)^*x(t;\omega)\,dt, u(s;\omega)\right]ds.$$

Let

$$w(t;\omega) = \int_t^1 B(t)^*\phi(t)^{*-1}\phi(s)^*x(s;\omega)\,ds.$$

Then observe that

$$E[w(t;\omega), u(t;\omega)] = E(E([w(t;\omega), u(t;\omega)]/\mathscr{B}_Z(t))),$$

and since $u(t;\omega)$ is measurable $\mathscr{B}_Z(t)$, we have that this is equal to

$$E([E((t;\omega)/\mathscr{B}_Z(t)), u(t;\omega)]).$$

Hence, it follows that L^*x is given by

$$L^*x = w, \qquad w(t;\omega) = B(t)^* \int_t^1 \phi(t)^{*-1}\phi(s)^*E(x(s;\omega)/\mathscr{B}_Z(t))\,ds. \tag{3.3}$$

In particular then we note that u_0 is given by

$$u_0(t;\omega) = (-1/\lambda)B(t)^* \int_t^1 \phi(t)^{*-1}\phi(s)^*Q(s)E(\hat{x}(s;\omega)/\mathscr{B}_Z(t))\,ds. \tag{3.4}$$

Next let us note that u_0 has also the alternate form:

$$u_0 = (L^*L + \lambda I)^{-1}L^*v, \tag{3.5}$$

where I is the identity operator, and it is readily verified that L^*w is actually in $\mathscr{H}_{\mathscr{L}}$, and that L^*L maps $\mathscr{H}_{\mathscr{L}}$ into itself. Hence it follows that u_0 is actually in $\mathscr{H}_{\mathscr{L}}$, and hence $\hat{x}(t;\omega)$ is actually Gaussian. Now, u_0 is unique, so that if we can determine $\hat{x}(t;\omega)$ as a Gaussian *Markov* process and still satisfy (3.4), we know that we would obtain the optimal control. But if $\hat{x}(t;\omega)$ were Markovian we must have

$$E(\hat{x}(s;\omega)/\mathscr{B}_Z(t)) = R(s;t)\hat{x}(t;\omega), \qquad t \leqq s, \tag{3.6}$$

and (3.4) would then yield that u_0 must have the form

$$u_0(t;\omega) = (-1/\lambda)B(t)^*P(t)\hat{x}(t;\omega). \tag{3.7}$$

But if we substitute this into (2.6a), we note that we must have

$$E(\hat{x}(s;\omega)/\mathscr{B}_Z(t)) = M(s)M(t)^{-1}\hat{x}(t;\omega), \qquad t \leqq s, \tag{3.8}$$

where $M(t)$ is a fundamental matrix solution of

$$\dot{M}(t) = (A(t) + (-1/\lambda)B(t)B(t)^*P(t))M(t),$$

and from (3.4) we must have

$$P(t) = \phi(t)^{*-1}\int_t^1 \phi(s)^*Q(s)M(s)\,ds M(t)^{-1}. \tag{3.9}$$

From this it is clear that $P(t)$ must satisfy

$$\dot{P}(t) + A(t)^*P(t) + Q(t) + P(t)(A(t) + (-1/\lambda)B(t)B(t)^*P(t)) = 0,$$
$$P(1) = 0. \tag{3.10}$$

But this is of course the familiar Riccati equation which is known to have a unique solution, and thus the optimal control is indeed given by (3.7), where $P(t)$ is the unique solution of (3.10). We have thus exploited the fact that the optimal solution is unique, and the only prescience we have used is to try Markov solution but surely this is very natural to try after a look at the form of (3.4).

REFERENCES

[1] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131–212.

[2] H. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.

[3] W. FLEMMING, *Optimal continuous parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–508.

[4] A. V. BALAKRISHNAN, *Optimal control problems in Banach space*, this Journal, 3 (1965), pp. 152–180.

[5] J. L. LIONS, *Contrôle optimale de systèmes gouvernés par des équations aux derivées partielles*, Dunod, Paris, 1969.

[6] T. KAILATH, *An innovations approach to least squares estimation*, IEEE Trans. Automatic Control, 13 (1968), pp. 646–655.

[7] A. V. BALAKRISHNAN, *A Martingale approach to linear recursive state estimation*, this Journal, 10 (1972).

[8] E. NELSON, *Dynamical Theories of Brownian Motion*, Princeton University Press, Princeton, N.J., 1967.

[9] A. V. BALAKRISHNAN, *Stochastic Differential Systems*, Academic Press, New York, to be published.

[10] K. ITÔ, *Multiple Wiener integral*, J. Math. Soc. Japan, 3 (1951), pp. 157–169.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

CONTROLLABILITY, OBSERVABILITY AND OPTIMAL FEEDBACK CONTROL OF AFFINE HEREDITARY DIFFERENTIAL SYSTEMS*

M. C. DELFOUR† AND S. K. MITTER‡

Abstract. This paper is concerned with two aspects of the control of affine hereditary differential systems. They are (i) the theory of various types of controllability and observability for such systems and (ii) the problem of optimal feedback control with a quadratic cost. The study is undertaken within the framework of hereditary differential systems with initial data in the space M^2 (cf. Delfour and Mitter [6], [7]). The main result of this paper is the existence and characterization of the optimal feedback operator for the system.

1. Introduction. Perhaps the most useful part of optimal control theory for ordinary differential equations is the theory of optimal control of linear differential systems with a quadratic cost criterion. This theory is also the most complete, both for systems evolving in a finite-time interval as well as over an infinite-time interval. It is well known that in the finite-time case the optimal control can be expressed in linear feedback form, where the "feedback gains" satisfy a matrix differential equation of Riccati type. In the infinite-time case by using the theory of controllability and observability, the asymptotic behavior of the controlled system can be studied and a rather complete solution to the problem is available.

The present paper is concerned with (i) generalization of the theory of controllability and observability to affine hereditary differential systems and (ii) a study of the optimal feedback control problem for affine hereditary differential systems with a quadratic cost. The theory is currently being completed in order to show the relation of the theory of controllability and observability to the infinite-time quadratic cost problem.

The optimal control problem studied in this paper was first formulated and studied by Krasovskii [23], [24] using the space of continuous functions as the space of initial data and using dynamic programming arguments. This problem has also been studied by Ross and Flügge-Lotz [30], Eller, Aggarwal and Banks [13], Kushner and Barnea [25] and Alekal, Brunovsky, Chyung and Lee [1], in each case using Carathéodory–Hamilton–Jacobi type arguments. The basic disadvantage of the method used by these authors is that it necessitates a direct study of a complicated set of coupled ordinary and first order partial differential equations before the existence of a feedback control can be asserted.

In Delfour and Mitter [6], [7] we have developed a theory of hereditary differential systems where the initial datum is chosen to lie in the space

* Received by the editors September 13, 1971, and in revised form January 21, 1972. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Centre de Recherches Mathématiques, Université de Montréal, Montréal 101, Québec, Canada.

‡ Electronic Systems Laboratory and Electrical Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The work of this author was supported in part by the National Science Foundation under Grant GK-25781 and by the Air Force Office of Scientific Research under Grant AFOSR-70-1941.

$M^p(-b,0;H)$, $1 \leqq p < \infty$. In particular, $M^2(-b,0;H)$ is a Hilbert space. By choosing the initial datum to lie in $M^2(-b,0;H)$, the techniques developed by J. L. Lions [26] for the control of parabolic partial differential equations can be appropriately modified to solve the optimal feedback control problem for affine hereditary differential systems. It should be emphasized that in contrast to the Hamilton–Jacobi method this is a direct method where the existence of the "feedback operator" is first demonstrated and it is then shown to satisfy an operational differential equation of Riccati type. Part of the results on the feedback control problem were announced in Delfour and Mitter [10].

The concepts of controllability and observability for hereditary differential systems are also studied within the framework developed in Delfour and Mitter [6], [7]. This is accomplished by using certain results on controllability and observability of abstract linear control systems (cf. Delfour and Mitter [11]). We present necessary and sufficient conditions for various types of controllability and we examine the dual system. We also show how various existing results on controllability fit into the framework adopted in this paper (cf. A. F. Buckalo [2], Chyung and Lee [3], D. R. Haley [19], Kirillova and Churakova [22], G. S. Tahim [32] and L. Weiss [33]–[36]).

1.1. Notation and terminology. Given two real linear spaces X and Y and a linear map $T: X \to Y$, the image of T in Y will be denoted by $\mathrm{Im}\,(T)$ and the kernel of T in X by $\mathrm{Ker}\,(T)$. Let H and K be two Hilbert spaces and $T: H \to K$ be a continuous linear map. The adjoint of T will be denoted $T^* (\in \mathscr{L}(K^*, H^*))$. When $H = K$ we shall say that T is self-adjoint if $T^* = T$ and we shall write $T \geqq 0$ for a positive self-adjoint operator ($(x|Tx) \geqq 0$ for all x) and $T > 0$ for a positive definite self-adjoint operator ($(x|Tx) > 0$ for all $x \neq 0$). The identity map in $\mathscr{L}(H)$ is written I. The restriction of the map $x: [0, \infty[\to X$ to the interval $[0, t]$ is denoted $\pi_t x$ for all $t \in]0, \infty[$. The set of real numbers is denoted by R.

In the sequel we shall abbreviate hereditary differential system as HDS.

2. Basic properties of affine HDS. Let H and U be Hilbert spaces. Let $N \geqq 1$ be an integer, let $a > 0, 0 = \theta_0 > \theta_1 > \cdots > \theta_N = -a$ be real numbers and $b \in [a, \infty]$. Let $I(\alpha, \beta) = R \cap [\alpha, \beta]$ for any $\alpha < \beta$ in $[-\infty, \infty]$. Let $|\cdot|_H$ (resp. $|\cdot|_U$) and $(\cdot|\cdot)_H$ (resp. $(\cdot|\cdot)_U$) denote the norm and inner products on H (resp. U).

2.1. Space of initial data and space of solutions. Our first task consists of choosing an appropriate space of initial data.

Consider the space $\mathscr{L}^2(-b,0;H)$ (not to be confused with $L^2(-b,0;H)$) of all maps $I(-b,0) \to H$ which are square integrable in $I(-b,0)$ endowed with the seminorm

$$\|y\|_{M^2} = \left[|y(0)|_H^2 + \int_{-b}^{0} |y(\theta)|_H^2 \, d\theta \right]^{1/2}.$$

The quotient space of $\mathscr{L}^2(-b,0;H)$ by the linear subspace of all y such that $\|y\|_{M^2} = 0$ is a Hilbert space which is isometrically isomorphic to the product space $H \times L^2(-b,0;H)$. It will be denoted by $M^2(-b,0;H)$ and its norm by $\|\cdot\|_{M^2}$. The isomorphism between $H \times L^2(-b,0;H)$ and $M^2(-b,0;H)$ is denoted by κ.

In order to discuss the Cauchy problem we must also describe the space in which solutions will be sought. Let $1 \leqq p < \infty$, $t_0 \in R$. For all $t \in]t_0, \infty[$ we denote by $AC^p(t_0, t; H)$ the vector space of all absolutely continuous maps $[t_0, t] \to H$ with a derivative in $L^p(t_0, t; H)$. When $AC^p(t_0, t; H)$ is endowed with the norm

$$\|x\|_{AC^p} = \left[|x(t_0)|_H^p + \int_{t_0}^t \left| \frac{dx}{ds}(s) \right|_H^p ds \right]^{1/p},$$

it is a Banach space isometrically isomorphic to $H \times L^p(t_0, t; H)$. In particular, $AC^2(t_0, t; H)$ is a Hilbert space. We shall also need $C(t_0, t; H)$, the Banach space of all continuous maps $[t_0, t] \to H$ endowed with the sup norm $\|\cdot\|_C$.

When we consider the evolution of a system in an infinite-time interval it is useful and quite natural to introduce the following spaces. Let $\pi_t(x)$ be the restriction of the map $x: [t_0, \infty[\to H$ to the interval $[t_0, t]$, $t \in]t_0, \infty[$. Denote by $L^p_{\text{loc}}(t_0, \infty; H)$, $AC^p_{\text{loc}}(t_0, \infty; H)$ and $C_{\text{loc}}(t_0, \infty; H)$ the vector space of all maps $x: [t_0, \infty[\to H$ such that for all $t \in]t_0, \infty[$, $\pi_t(x)$ is in $L^p(t_0, t; H)$, $AC^p(t_0, t; H)$ and $C(t_0, t; H)$, respectively. They are Fréchet spaces (cf. Delfour [5]) when their respective topologies are defined by the saturated family of seminorms $q_t(x) = \|\pi_t(x)\|_F$, $t \in]t_0, \infty[$, where F is either L^p, AC^p or C.

2.2. System description. Consider the affine hereditary differential system $\mathscr{A}$ defined on $[0, \infty[$:

$$\begin{aligned} \frac{dx}{dt}(t) = {} & A_{00}(t)x(t) + \sum_{i=1}^{N} A_i(t) \begin{Bmatrix} x(t+\theta_i), & t+\theta_i \geqq 0 \\ h(t+\theta_i), & t+\theta_i < 0 \end{Bmatrix} \\ & + \int_{-b}^{0} A_{01}(t, \theta) \begin{Bmatrix} x(t+\theta), & t+\theta \geqq 0 \\ h(t+\theta), & t+\theta < 0 \end{Bmatrix} d\theta \\ & + B(t)v(t) + f(t) \quad \text{a.e. in } [0, \infty), \\ x(0) = {} & h(0), \quad h \in M^2(-b, 0; H), \end{aligned} \tag{2.1}$$

where A_{00} and A_i $(i = 1, 2, \cdots, N)$ are in $L^\infty_{\text{loc}}(0, \infty; \mathscr{L}(H))$, $A_{01} \in L^\infty_{\text{loc}}(0, \infty; -b, 0; \mathscr{L}(H))$, $B \in L^\infty_{\text{loc}}(0, \infty; \mathscr{L}(U, H))$, $v \in L^2_{\text{loc}}(0, \infty; U)$ and $f \in L^2_{\text{loc}}(0, \infty; H)$.

v is to be thought of as the control to be applied to the system and f is a known external input to the system. Under the above hypotheses, (2.1) has a unique solution $\phi(\cdot; h, v)$ in $AC^2_{\text{loc}}(0, \infty; H)$ and the map

$$(h, v) \mapsto \phi(\cdot; h, v): M^2(-b, 0; H) \times L^2_{\text{loc}}(0, \infty; U) \to AC^2_{\text{loc}}(0, \infty; H) \tag{2.2}$$

is affine and continuous (cf. Delfour and Mitter [6], [7] and Delfour [5]). We also have the variation of constants formula

$$\phi(t; h, v) = \Phi(t, 0)h + \int_0^t \Phi^0(t, s)B(s)v(s)\, ds + \int_0^t \Phi^0(t, s)f(s)\, ds, \tag{2.3}$$

where

$$\Phi(t, s)h = \Phi^0(t, s)h(0) + \int_{-b}^{0} \Phi^1(t, s, \alpha)h(\alpha)\, d\alpha,$$

and $\Phi^0(t, s) \in \mathscr{L}(H)$ is the unique solution in $AC^2_{\text{loc}}(s, \infty; \mathscr{L}(H))$ of the system

$$\begin{aligned}\frac{\partial \Phi^0}{\partial t}(t, s) &= A_{00}(t)\Phi^0(t, s) + \sum_{i=1}^{N} A_i(t)\begin{cases}\Phi^0(t + \theta_i, s), & t + \theta_i \geqq s\\ 0, & \text{otherwise}\end{cases}\\ &\quad + \int_{-b}^{0} A_{01}(t, \theta)\begin{cases}\Phi^0(t + \theta, s), & t + \theta \geqq s\\ 0, & \text{otherwise}\end{cases} d\theta \quad \text{a.e. in } [s, \infty[\end{aligned} \tag{2.4}$$

and

$$\begin{aligned}\Phi^1(t, s, \alpha) &= \sum_{i=1}^{N} \begin{cases}\Phi^0(t, s + \alpha - \theta_i)A_i(s + \alpha - \theta_i), & \alpha + s - t < \theta_i \leqq \alpha\\ 0, & \text{otherwise}\end{cases}\\ &\quad + \begin{cases}\displaystyle\int_{-b}^{\alpha} \Phi^0(t, s + \alpha - \theta)A_{01}(s + \alpha - \theta, \theta)\, d\theta, & s + \alpha \leqq t - b\\ \displaystyle\int_{\alpha - t + s}^{\alpha} \Phi^0(t, s + \alpha - \theta)A_{01}(s + \alpha - \theta, \theta)\, d\theta, & s + \alpha > t - b\end{cases}.\end{aligned} \tag{2.5}$$

2.3. State equation of the system.

DEFINITION 2.1. Let $f = 0$, $v = 0$ in (2.1). The *evolution of the state* of the homogeneous system is given by the map

$$t \mapsto \tilde{\phi}(t; h) : [0, \infty[\to M^2(-b, 0; H) \tag{2.6}$$

defined as

$$\tilde{\phi}(t; h)(\theta) = \begin{cases}\phi(t + \theta; h), & t + \theta \geqq 0,\\ h(t + \theta), & t + \theta < 0.\end{cases} \tag{2.7}$$

It is easy to verify the following theorem.

THEOREM 2.2. *Consider* (2.1) *with* $f = 0, v = 0$ *on* $[s, \infty[$ *with initial datum* h *at time* s. *Let* $\tilde{\phi}_s(\cdot\,; h)$ *denote the solution of this system in* $AC^2_{\text{loc}}(s, \infty; H)$. *The map* $(t, s) \mapsto \tilde{\phi}_s(t; h)$ *generates a two-parameter semigroup* $\tilde{\Phi}(t, s)$ *satisfying the following properties*:

(i) $\tilde{\Phi}(t, s) \in \mathscr{L}(M^2), t \geqq s \geqq 0$;

(ii) $\tilde{\Phi}(t, r) = \tilde{\Phi}(t, s)\tilde{\Phi}(s, r), t \geqq s \geqq r \geqq 0$;

(iii) $t \mapsto \tilde{\Phi}(t, s)h : [s, \infty[\to M^2$ *is continuous for all* $h \in M^2$ *and* $s \in [0, \infty[$;

(iv) $\tilde{\Phi}(s, s) = I$, *where* I *is the identity operator in* $\mathscr{L}(M^2)$;

(v) *for* $t - s \geqq b$, $\tilde{\Phi}(t, s) : M^2 \to M^2$ *is compact* (*i.e., maps bounded sets into relatively compact sets*);

(vi) *Let* $\mathscr{D} = AC^2(-b, 0; H) \cap M^2(-b, 0; H)$. *Then for all* $h \in \mathscr{D}$, $\tilde{\Phi}(t, s)h \in \mathscr{D}$. ▨

Since M^2 is isomorphic to $H \times L^2(-b, 0; H)$, $\tilde{\Phi}(t, s)$ can be decomposed into two operators $\tilde{\Phi}^0(t, s) \in \mathscr{L}(H, M^2)$ and $\tilde{\Phi}^1(t, s) \in \mathscr{L}(L^2(-b, 0; H), M^2)$ such that

$$\tilde{\Phi}(t, s)h = \tilde{\Phi}^0(t, s)h^0 + \tilde{\Phi}^1(t, s)h^1,$$

where

$$[\tilde{\Phi}^0(t, s)h^0](\alpha) = \begin{cases}\Phi^0(t + \alpha, s)h^0, & t + \alpha \geqq s,\\ 0, & t + \alpha < s,\end{cases} \tag{2.8}$$

and

$$[\tilde{\Phi}^1(t,s)h^1](\alpha) = \begin{cases} \int_{-b}^{0} \Phi^1(t+\alpha, s, \eta)h^1(\eta)\,d\eta, & t+\alpha \geqq s, \\ h^1(t+\alpha-s), & t+\alpha < s. \end{cases} \tag{2.9}$$

Finally corresponding to (2.1) we have the *state equation* in *integral form*

$$\tilde{\phi}(t;h,v) = \tilde{\Phi}(t,0)h + \int_0^t \tilde{\Phi}^0(t,s)B(s)v(s)\,ds + \int_0^t \tilde{\Phi}^0(t,s)f(s)\,ds. \tag{2.10}$$

We now wish to obtain the state equation in differential form. We first construct an unbounded operator $\tilde{A}(t)$ whose domain is

$$\mathscr{D} = AC^2(-b,0;H) \cap M^2(-b,0;H).$$

For this purpose define the linear maps

$$\tilde{A}^0(t):\mathscr{D} \to H \quad \text{and} \quad \tilde{A}^1:\mathscr{D} \to L^2(-b,0;H)$$

as follows:

$$\tilde{A}^0(t)h = A_{00}(t)h(0) + \sum_{i=1}^{N} A_i(t)h(\theta_i) + \int_{-b}^{0} A_{01}(t,\theta)h(\theta)\,d\theta \tag{2.11}$$

and

$$(\tilde{A}^1 h)(\theta) = \frac{dh(\theta)}{d\theta}. \tag{2.12}$$

From the operators $\tilde{A}^0(t)$ and $\tilde{A}^1$ we construct the unbounded operator $\tilde{A}(t):\mathscr{D} \to M^2(-b,0;H)$ as

$$[\tilde{A}(t)h](\alpha) = \begin{cases} \tilde{A}^0(t)h, & \alpha = 0, \\ [\tilde{A}^1 h](\alpha), & \alpha \neq 0. \end{cases} \tag{2.13}$$

Define also the operator $\tilde{B}(t):U \to M^2(-b,0;H)$ as

$$[\tilde{B}(t)u](\alpha) = \begin{cases} B(t)u, & \alpha = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2.14}$$

and $\tilde{f}(t) \in M^2(-b,0;H)$ as

$$[\tilde{f}(t)](\alpha) = \begin{cases} f(t), & \alpha = 0, \\ 0, & \text{otherwise}. \end{cases} \tag{2.15}$$

We then have the following theorem.

THEOREM 2.3. (i) *For all $h \in \mathscr{D}$ and all $u \in L^2_{\text{loc}}(0,\infty;U)$, the system*

$$\begin{aligned} &\frac{dy}{dt}(t) = \tilde{A}(t)y(t) + \tilde{B}(t)u(t) + \tilde{f}(t) \quad \text{a.e. in } [0,\infty[, \\ &y(0) = h \end{aligned} \tag{2.16}$$

has a unique solution in $AC^2_{\text{loc}}(0,\infty;M^2)$ which coincides with $\tilde{\phi}(\cdot\,;h,u)$.

(ii) *The map $(h,u) \mapsto \Lambda(h,u) = \tilde{\phi}(\cdot\,;h,u):\mathscr{D} \times L^2_{\text{loc}}(0,\infty;U) \to AC^2_{\text{loc}}(0,\infty;M^2)$ can be lifted to a unique continuous affine map $\tilde{\Lambda}:M^2 \times L^2_{\text{loc}}(0,\infty;U) \to C_{\text{loc}}(0,\infty;M^2)$ and for all pairs (h,u), $\tilde{\Lambda}(h,u)$ coincides with $\tilde{\phi}(\cdot\,;h,u)$.*

Proof. Cf. Delfour and Mitter [6], [7]. ▨

Remark 2.4. In the autonomous case ($\tilde{A}(t) = \tilde{A} = \text{const.}$) the semigroup $\{\tilde{\Phi}(t,s)\}$ becomes a one-parameter semigroup $\{\tilde{\Phi}(t)\}$ and its infinitesimal generator is precisely $\tilde{A}$ and the domain of $\tilde{A}$ is $\mathscr{D}$ (cf. Hale [16] for analogous considerations).

2.4. Adjoint systems. One of the truly fascinating aspects of linear HDS is the existence of two types of adjoint systems: a *hereditary adjoint system* and a *topological adjoint system.*

2.4.1. Hereditary adjoint system. The *hereditary adjoint system* is defined in the interval $[0, T]$ for some $T > 0$:

$$\begin{aligned}\frac{dp}{dt}(t) + A_{00}(t)^*p(t) + \sum_{i=1}^{N} \begin{cases} A_i(t-\theta_i)^*p(t-\theta_i), & t-\theta_i \leqq T \\ 0, & t-\theta_i > T \end{cases} \\ + \int_{-b}^{0} \begin{cases} A_{01}(t-\theta)^*p(t-\theta), & t-\theta \leqq T \\ 0, & t-\theta > T \end{cases} d\theta + g(t) = 0 \quad \text{a.e. in } [0, T],\end{aligned} \tag{2.17}$$

$$p(T) = k^0, \quad k^0 \in H, \tag{2.18}$$

for some $g \in L^2(0, T; H)$. Under the hypotheses of § 2.2, (2.17)–(2.18) has a unique solution $\psi(\cdot\,; T, k^0)$ in $AC^2(0, T; H)$ and the map

$$k^0 \mapsto \psi(\cdot\,; T, k^0): H \to AC^2(0, T; H) \tag{2.19}$$

is affine and continuous. We also have the *variation of constants* formula

$$\psi(t; T, k^0) = \Phi^0(T, t)^*k^0 + \int_t^T \Phi^0(r, t)^*g(r)\, dr, \tag{2.20}$$

where Φ^0 is defined in (2.4) (cf. Delfour and Mitter [7], [9]).

It will be convenient to construct the following unbounded operator $\tilde{A}_T^0(t)$: $\mathscr{D}^* \to H$ with appropriate domain $\mathscr{D}^*$:

$$\begin{aligned}\tilde{A}_T^0(t)h = A_{00}(t)^*h(0) + \sum_{i=1}^{N} \begin{cases} A_i(t-\theta_i)^*h(\theta_i), & t-\theta_i \leqq T \\ 0, & t-\theta_i > T \end{cases} \\ + \int_{-b}^{0} \begin{cases} A_{01}(t-\theta, \theta)^*h(\theta), & t-\theta \leqq T \\ 0, & t-\theta > T \end{cases} d\theta.\end{aligned} \tag{2.21}$$

Equation (2.19) can now be rewritten in a condensed form:

$$\begin{aligned}&\frac{dp}{dt}(t) + \tilde{A}_T^0(t)p_t + g(t) = 0 \quad \text{a.e. in } [0, T], \\ &p(T) = k^0 \in H,\end{aligned} \tag{2.22}$$

where $p_t \in M^2$ is defined as

$$p_t(\theta) = \begin{cases} p(t-\theta), & t-\theta \leqq T, \\ 0, & \text{otherwise.} \end{cases} \tag{2.23}$$

Remark 2.5. Even when $A_{00}, A_1, \cdots, A_N$ and A_{01} are time invariant, the hereditary adjoint system depends on both T and the time t.

2.4.2. Topological adjoint system. Owing to some delicate technical considerations, we restrict our attention to the autonomous case ($\tilde{A}(t) = \tilde{A} = \text{const.}$). $\tilde{A}^*$ denotes the M^2-adjoint of $\tilde{A}$.

THEOREM 2.6. *Given $T > 0$, the densely defined closed operator $-\tilde{A}^*$ generates the one-parameter semigroup $\{\Psi(T-t)^*\}$ and for all $k \in \mathscr{D}(\tilde{A}^*)$, $z(t) = \Psi(T-t)^*k$*

is the unique solution in $AC^2(0, T; M^2)$ *of the equation*

$$\begin{aligned} &\frac{dz}{dt}(t) + \tilde{A}^* z(t) = 0 \quad \text{in } [0, T], \\ &z(T) = k \in \mathscr{D}(A^*). \end{aligned} \tag{2.24}$$

Proof. Cf. Delfour and Mitter [7]. ▨

For obvious reasons system (2.24) will be referred to as the *topological adjoint system.*

3. Controllability and observability. In §§ 3.1 and 3.2 we successively look at various notions of controllability and observability, discuss their relative merits and prove various results on controllability and observability. In the last section we construct a system which is dual to the original system. The relationship between controllability and observability and the feedback problem will be considered in a forthcoming paper.

3.1. Controllability. The notions of controllability for hereditary differential systems have been explored by several authors since 1965 and precise conditions have been presented for controllability (cf. G. S. Tahim [32], Chyung and Lee [3], Kirillova and Churakova [22], L. Weiss [33]–[36], A. F. Buckalo [2] and A. Halanay [17]) of different types. In this section we look at two types of controllability in the framework of the space M^2, derive necessary and sufficient conditions and discuss the relationship of earlier results in the literature with ours.

DEFINITION 3.1. (i) The data $h \in M^2$ is *controllable* (resp. M^2-*controllable*) *at time* T to $x \in H$ (resp. $k \in M^2$) if there exists a sequence $\{u_n\}$ in $L^2(0, T; U)$ such that $\phi(T; h, u_n)$ (resp. $\tilde{\phi}(T; h, u_n)$) converges to x (resp. k). System $\mathscr{A}$ is *controllable* (resp. M^2-*controllable*) *at time* T if all $h \in M^2$ are controllable (resp. M^2-controllable) at time T to all $x \in H$ (resp. $k \in M^2$).

(ii) The data $h \in M^2$ is *controllable to the origin* (resp. *to the zero function*) if there exists a finite time $T > 0$ for which h is controllable to $0 \in H$ (resp. $0 \in M^2$) at time T. If every h in M^2 is *controllable to the origin* (resp. *to the zero function*), system $\mathscr{A}$ is said to be *controllable to the origin* (resp. *to the zero function*).

DEFINITION 3.2. (i) The data $h \in M^2$ is *strictly controllable* (resp. M^2-*controllable*) *at time* T to $x \in X$ (resp. $k \in M^2$) if there exists u in $L^2(0, T; U)$ such that $\phi(T; h, u) = x$ (resp. $\tilde{\phi}(T; h, u) = k$). System $\mathscr{A}$ is said to be *strictly controllable* (resp. M^2-*controllable*) *at time* T if every $h \in M^2$ is strictly controllable (resp. M^2-controllable) at time T to all $x \in H$ (resp. $k \in M^2$).

(ii) The data $h \in M^2$ is *strictly controllable to the origin* (resp. *to the zero function*) if there exists a finite time $T > 0$ for which h is strictly controllable to $0 \in H$ (resp. $0 \in M^2$) at time T. If all data h in M^2 are strictly controllable to the origin (resp. to the zero function), system $\mathscr{A}$ is said to be *controllable to the origin* (resp. *to the zero function*).

For completeness we have included this last definition.

DEFINITION 3.3. Let $\mathscr{H}$ be a linear subspace of M^2. System $\mathscr{A}$ is *strictly controllable to a function* ψ *in* $\mathscr{H}$ if for each $h \in M^2$, there exist a finite time $T > 0$ and a control map $u \in L^2(0, T; U)$ such that $\tilde{\phi}(T; h, u) = \psi$. System $\mathscr{A}$ is said to be *strictly controllable to the space* $\mathscr{H}$ if it is strictly controllable to all functions ψ of $\mathscr{H}$.

PROPOSITION 3.4. *When H is finite-dimensional the notion of strict controllability at time T (resp. strict controllability to the origin) is equivalent to the notion of controllability at time T (resp. controllability to the origin).*

In this paper we shall not consider the "strict" notions unless we are in the situation of Proposition 3.4. The following results are obtained directly from the definitions.

PROPOSITION 3.5. (i) *$\mathscr{A}$ is never strictly M^2-controllable at time T.*

(ii) *For all $T < b$, $\mathscr{A}$ is never M^2-controllable at time T or controllable to the zero function.*

(iii) *The controllability of $\mathscr{A}$ at time T is a necessary condition for the M^2-controllability of $\mathscr{A}$ at time T.* ▧

Remark. (i) When $b = \infty$, $\mathscr{A}$ is never controllable to the zero function or M^2-controllable at any finite $T \geqq 0$.

(ii) Proposition 3.5 (i) implies that when $\mathscr{A}$ is M^2-controllable at time T, the initial states in M^2 are only strictly controllable to points in a dense subspace of M^2 which is different from M^2.

PROPOSITION 3.6. *$\mathscr{A}$ is controllable to the origin (to the zero function) if there exists a finite time $T > 0$ such that $\mathscr{A}$ is controllable (M^2-controllable) at time T.* ▧

Remark. A similar statement is true for the "strict" notions.

All the above definitions were originally given in the literature for initial data in $C(-b, 0; H)$ rather than $M^2(-b, 0; H)$. As for the space of control maps, it is safer and technically more advantageous to use the larger $L^2(0, T; U)$ rather than $L^\infty(0, T; U)$ or the space of piecewise continuous maps. Whether the control map can be picked in a smaller subspace of $L^2(0, T; U)$ will depend on the nature of the operators A_{00}, A_i $(i = 1, \cdots, N)$ and A_{01}. If they are "sufficiently nice" so will the control maps be. Table 1 summarizes some of the details concerning previous research. In all cases, $H = R^n$ and the controllability is strict (s.c.).

TABLE 1

	Types of Controllability	Control Maps
Chyung and Lee, 1966 [3]	• $A_{01} = 0$, continuous matrices A_{00}, A_i $(i = 1, \cdots, N)$ and B. • s.c. at time T.	$L^\infty(0, T; H)$
Kirillova and Churakova, 1967 [22]	• $N = 1$, $A_{01} = 0$, constant matrices A_{00}, A_1 and B. • s.c. to the origin and to the zero function.	piecewise continuous maps $[0, \infty[\to U$
L. Weiss, 1967 [33] and 1970 [34]	• $N = 1$, $A_{01} = 0$, continuous matrices A_{00}, A_1 and B. • s.c. to the origin, to the zero function and to a function.	$L^\infty_{1\infty}(0, \infty; U)$
A. F. Buckalo, 1968 [2]	• $N = 1$, $A_{01} = 0$, $n - 1$ differentiable matrices A_{00}, A_1 and B. • s.c. to the zero function.	$L^\infty_{1\infty}(0, \infty; U)$
A. Halanay, 1970 [18]	• $N = 1$, $A_{01} = 0$, special constant matrices A_{00}, A_1 and B. • "complete controllability" (=s.c. to the zero function for all $T > a$).	piecewise continuous maps $[0, \infty[\to U$

Definitions 3.1 (ii), (3.2) (ii) and 3.3 are conceptually interesting but technically difficult to deal with since the final time T is not fixed. Even from the engineering standpoint it is desirable to have a uniform bound on T independent of the initial data h in M^2. In fact, most conditions for "controllability of Definitions 3.1 (ii), 3.2 (ii) and 3.3" are only sufficient. They make use of Proposition 3.6, the converse of which is obviously not true. The notion of M^2-controllability is new in the context of hereditary differential systems, though the idea of density has often been used in partial differential equations where it naturally arises. It is clear that at time $t > 0$ the state $\tilde{\phi}(t; h, u)$ will be absolutely continuous in $[-t, 0]$ (see (2.7)). Thus it will be impossible to synthesize an M^2-map or even a continuous map defined in $I(-b, 0)$ which is not at least differentiable in the interval $[-t, 0]$. For all the above reasons we shall limit the scope of our investigation to the notions of controllability of Definitions 3.1 (i) and 3.2 (i).

THEOREM 3.7. *The following statements are equivalent*:

(i) $\mathscr{A}$ *is controllable (resp.* M^2*-controllable) at time* T;

(ii) *the map* $u \mapsto S(T)u = \int_0^T \Phi^0(T, s)B(s)u(s)\,ds : L^2(0, T; U) \mapsto H$ *(resp.* $u \mapsto \tilde{S}(T)u = \int_0^T \tilde{\Phi}^0(T, s)B(s)u(s)\,ds : L^2(0, T; U) \mapsto M^2(-b, 0; H)$*) has a dense image in* H *(resp.* $M^2(-b, 0; H)$*)*;

(iii) *the map* $x \mapsto S(T)^*x : H \to L^2(0, T; U)$ *(resp.* $k \mapsto \tilde{S}(T)^*k : M^2(-b, 0; H) \to L^2(0, T; U)$*), where* $(S(T)^*x)(t) = B(t)^*\Phi^0(T, t)^*x$ *(resp.* $(\tilde{S}(T)^*k)(t) = B(t)^*\tilde{\Phi}^0(T, t)^*k$*), is injective*;

(iv) *the symmetric operator*

$$W_c(T) = \int_0^T \Phi^0(T, s)B(s)B(s)^*\Phi^0(T, s)^*\,ds$$

$$(\textit{resp. } \tilde{W}_c(T) = \int_0^T \tilde{\Phi}^0(T, s)B(s)B(s)^*\tilde{\Phi}^0(T, s)^*\,ds)$$

is positive definite.

Proof. This is a corollary to Delfour and Mitter [11, Thm. 9 and Cor. 10]. ▧

COROLLARY 3.8. *Let* $H = R^n$. (i) *The condition*

$$\operatorname{rank}(W_c(T)) = n \tag{3.1}$$

is necessary and sufficient for the strict controllability of $\mathscr{A}$ *at time* T.

(ii) *Condition* (3.1) *is necessary for the* M^2*-controllability of* $\mathscr{A}$ *at time* T.

(iii) *If there exists a time* T, $0 \leqq T < \infty$, *for which condition* (3.1) *holds, then system* $\mathscr{A}$ *is strictly controllable to the origin.* ▧

Remark. Part (i) of the corollary is due to Chyung and Lee [3] and part (iii) to L. Weiss [33, Lemma 1].

PROPOSITION 3.9. *Assume there exists* T, $0 < T < \infty$, *such that* $\operatorname{Im}(\Phi(T, 0)) = H$. *If all initial states* $h \in M^2$ *are controllable to the origin at time* T, *system* $\mathscr{A}$ *is controllable at time* T.

Proof. For all $h \in M^2$, $x \in H$ there exists $k \in M^2$ such that

$$-x + \Phi(T, 0)h = \Phi(T, 0)k. \tag{3.2}$$

Since $\mathscr{A}$ is controllable to the origin at time T there exists $\{u_n\}$ in $L^2(0, T; U)$ such that

$$\Phi(T, 0)k + \int_0^T \Phi^0(T, s)B(s)u_n(s)\,ds + \int_0^T \Phi^0(T, s)f(s)\,ds \to 0. \tag{3.3}$$

Hence $\mathscr{A}$ is controllable at time T by combining (3.2) and (3.3). ▨

Remark. The condition $\operatorname{Im}(\Phi(T,0)) = H$ is equivalent to have the "force free attainable set"

$$\{\Phi(T,0)h | h \in M^2\}$$

equal to H. When the property is true for all $T \geqq 0$, then system $\mathscr{A}$ is said to be *pointwise complete.* This definition is due to L. Weiss [33] who conjectured that for $H = R^n$ all systems of the form

$$\frac{dx}{dt}(t) = A_0x(t) + A_1x(t-a) + Bu(t), \quad t \geqq 0,$$

$$x(s) = h(s), \quad s \in [-a, 0],$$

are pointwise complete. This point has been investigated by V. M. Popov [29] who has shown that the conjecture is false for $n > 2$. Popov has further found necessary and sufficient conditions for the system to be pointwise complete.

Finally we restrict our attention to $H = R^n$ and systems of the form

$$\begin{aligned} &\frac{dx}{dt}(t) = A_0(t)x(t) + A_1(t)x(t-a) + B(t)u(t) + f(t), \quad t \geqq 0, \\ &x(s) = h(s), \quad s \in [-a, 0], \end{aligned} \tag{3.4}$$

where A_0 and A_1 are in $L^\infty_{\mathrm{loc}}(0, \infty; \mathscr{L}(R^n))$, $B \in L^\infty_{\mathrm{loc}}(0, \infty; \mathscr{L}(R^m, R^n))$ and $f \in L^2(0, \infty; R^n)$. Notice that

$$\begin{aligned} &\frac{\partial \Phi^0}{\partial t}(t,0) = A_0(t)\Phi^0(t,0), \quad t \in [0, a], \\ &\Phi^0(0,0) = \mathrm{I} \end{aligned} \tag{3.5}$$

and that for $T \geqq a$,

$$\begin{aligned} &\frac{\partial \Phi^0}{\partial s}(T,s) + \Phi^0(T,s)A_0(s) = 0, \quad s \in [T-a, T], \\ &\Phi^0(T,T) = I. \end{aligned} \tag{3.6}$$

This means that the force free attainable set is equal to R^n in the interval $[0, a]$ since

$$R^n = \operatorname{Im}(\Phi^0(t,0)) \subset \operatorname{Im}(\Phi(t,0)) \subset R^n.$$

Also, we have the following proposition.

PROPOSITION 3.10. (i) *If there exists* $T_0 \in [T-a, T[\cap [0, T]$ *for which the system*

$$\frac{dx}{dt}(t) = A_0(t)x(t) + B(t)u(t), \quad t \in [T_0, T],$$

is strictly controllable at time T, then system $\mathscr{A}$ *is strictly controllable at time T.*

(ii) *If in addition* A_0 *and B are respectively* $n-2$ *and* $n-1$ *times continuously differentiable in* $[T_0, T]$, *we can construct the* controllability matrix *of Silverman*

and Meadows [31]

$$Q_c(t) = [P_0(t) \vdots P_1(t) \vdots \cdots \vdots P_{n-1}(t)],$$

where

$$P_{k+1}(t) = -A_0(t)P_k(t) + \dot{P}_k(t),$$
$$P_0(t) = B(t),$$

and the condition of part (i) *is equivalent to the existence of some* $t \in [T_0, T]$ *for which* rank $Q_c(t) = n$. ▧

Remark. A. F. Buckalo [2] used the condition of L. Weiss and incorporated the ideas of Silverman and Meadows [31] to essentially obtain part (ii) of the above proposition. The classical rank condition is obtained when the matrices A_0 and B are not time dependent.

In addition to the above results one should mention the work of Kirillova and Churakova [22] and L. Weiss [34]. It is the first attempt to obtain direct conditions on the various matrices in contrast to the above results where the (strict) controllability of a nonhereditary system serves as a sufficient condition for the (strict) controllability of $\mathscr{A}$.

3.2. Observability. To our knowledge the notion of observability for HDS has not been studied in the published literature. We have seen that there are several notions of controllability. Likewise, there are more than one way to observe system $\mathscr{A}$ and different things to observe.

Let Y be a Hilbert space which might be thought of as the *observation space.* We can observe the map $\phi(\cdot\,; h, u)$ with an *observer* $Z \in L^\infty(0, T; \mathscr{L}(H, Y))$; the *observation* at time t is defined by

$$z(t; h, u) = Z(t)\phi(t; h, u). \tag{3.7}$$

We can also observe the map $\tilde{\phi}(\cdot\,; h, u)$ with an M^2*-observer* $\tilde{Z} \in L^\infty(0, T; \mathscr{L}(M^2, Y))$; the M^2*-observation* at time t is defined by

$$\tilde{z}(t; h, u) = \tilde{Z}(t)\tilde{\phi}(t; h, u). \tag{3.8}$$

Since $M^2(-b, 0; H)$ is isomorphic to $H \times L^2(-b, 0; H)$, there exist $\tilde{Z}^0(t) \in \mathscr{L}(H, Y)$ and $\tilde{Z}^1(t) \in \mathscr{L}(L^2(-b, 0; H), Y)$ such that

$$\tilde{Z}(t)(\kappa^{-1}(h^0, 0)) = \tilde{Z}^0(t)h^0 \tag{3.9}$$

and

$$\tilde{Z}(t)(\kappa^{-1}(0, h^1)) = \tilde{Z}^1(t)h^1. \tag{3.10}$$

Notice that our observer satisfies hypothesis (ii) in Definition 12 (cf. Delfour and Mitter [11]). Now starting from either of the above two types of observations, we can either determine the state $h \in M^2(-b, 0; H)$ or simply $h^0 \in H$.

DEFINITION 3.11. (i) System $\mathscr{A}$ is *observable* in $[0, T]$ if for all $h \in M^2(-b, 0; H)$ and $u \in L^2(0, T; U)$ the point $h^0 \in H$ can be uniquely determined from a knowledge of u, h^1 and the *observation map* $z(\cdot\,; h, u)$, where $\kappa(h) = (h^0, h^1)$ and κ is the isometric isomorphism between $M^2(-b, 0; H)$ and $H \times L^2(-b, 0; H)$.

(ii) System $\mathscr{A}$ is *strongly observable* in $[0, T]$ if for all $h \in M^2(-b, 0; H)$ and $u \in L^2(0, T; U)$, the state h can be uniquely determined from a knowledge of u and the *observation map* $z(\cdot\,; h, u)$.

(iii) System $\mathscr{A}$ is *M^2-observable* in $[0, T]$ if for all $h \in M^2(-b, 0; H)$ and $u \in L^2(0, T; U)$ the state h can be uniquely determined from u and the *observation map* $\tilde{z}(\cdot\,; h, u)$.

PROPOSITION 3.12. *Let* $\tilde{Z}^0(t) = Z(t)$.

(i) *$\mathscr{A}$ strongly observable $\Rightarrow$ $\mathscr{A}$ M^2-observable and observable.*

(ii) *For all $T < b$, $\mathscr{A}$ is not strongly observable in* $[0, T]$.

Proof. The proof follows from the definitions. ▧

Remarks. (i) When $b = \infty$, system $\mathscr{A}$ is never strongly observable in $[0, T]$ for all finite T.

(ii) When $\tilde{Z}^1(t) = 0$, strong observability and M^2-observability are equivalent.

PROPOSITION 3.13. *The following statements are equivalent*:

(i) *$\mathscr{A}$ is observable in* $[0, T]$;

(ii) *the map* $F^0 : H \to L^2(0, T; Y)$, *where* $((F^0)h^0)(t) = Z(t)\Phi^0(t, 0)h^0$, *is injective*;

(iii) *the map* $y \mapsto (F^0)^* y = \int_0^T \Phi^0(t, 0)^* Z(t)^* y(t)\, dt : L^2(0, T; Y) \to H$ *has a dense image in* H;

(iv) *the symmetric operator*

$$W_0^0(T) = \int_0^T \Phi^0(t, 0)^* Z(t)^* Z(t) \Phi^0(t, 0)\, dt \tag{3.11}$$

is positive definite.

Proof. The proof is similar to the proof of Theorem 3.7. ▧

COROLLARY 3.14. *Let* $H = R^n$. (i) *The condition*

$$\operatorname{rank}(W_0^0(T)) = n \tag{3.12}$$

is necessary and sufficient for the observability of $\mathscr{A}$ in $[0, T]$.

(ii) *Condition* (3.12) *is necessary for strong observability of system $\mathscr{A}$.*

Proof. The proof is similar to the proof of Corollary 3.8. ▧

COROLLARY 3.15. *Consider the system of equations* (3.4) *with observer* $Z \in L^\infty(0, T; \mathscr{L}(R^n, Y))$.

(i) *If there exists* $T_f \in [0, T] \cap [0, a]$ *for which the system*

$$\frac{dx}{dt}(t) + A_0(t)^* x(t) + Z(t)^* y(t) = 0 \quad \text{in } [0, T_f],$$

$$x(T_f) = h^0$$

is controllable at time 0, *then system $\mathscr{A}$ is observable in* $[0, T]$.

(ii) *If in addition A_0 and Z are respectively $n-2$ and $n-1$ times continuously differentiable in* $[0, T_f]$, *we can construct the* observability matrix *of Silverman and Meadows* [31],

$$Q_0(t) = [S_0(t) \vdots S_1(t) \vdots \cdots \vdots S_{n-1}(t)], \tag{3.13}$$

where

$$S_{k+1}(t) = A_0(t)^* S_k(t) + \dot{S}_k(t), \quad S_0(t) = Z(t)^*, \tag{3.14}$$

and the condition of part (i) *is equivalent to the existence of some* $t \in [0, T_f]$ *for which* rank $(Q_0(t)) = n$. ▧

Remark. The classical rank condition is obtained when the matrices A_0 and Z are not time dependent.

PROPOSITION 3.16. *The following statements are equivalent*:

(i) $\mathscr{A}$ *is strongly observable (resp.* M^2*-observable) in* $[0, T]$;

(ii) *the map* F *(resp.* $\tilde{F}$*)*: $M^2(-b, 0; H) \to L^2(0, T; Y)$ *defined by* $(Fh)(t) = Z(t)\Phi(t, 0)h$ *(resp.* $(\tilde{F}h)(t) = \tilde{Z}(t)\tilde{\Phi}(t, 0)h$*) is injective*;

(iii) *the map* F^* *(resp.* $\tilde{F}^*$*)*: $L^2(0, T; Y) \to M^2(-b, 0; H)$ *defined by*

$$F^*y = \int_0^T \Phi(t, 0)^* Z(t)^* y(t)\, dt \quad \left(\text{resp. } \tilde{F}^*y = \int_0^T \tilde{\Phi}(t, 0)^* \tilde{Z}(t)^* y(t)\, dt\right)$$

has a dense image in $M^2(-b, 0; H)$;

(iv) *the symmetric operator*

$$W_0(T) = \int_0^T \Phi(t, 0)^* Z(t)^* Z(t) \Phi(t, 0)\, dt$$

$$\left(\text{resp. } \tilde{W}_0(T) = \int_0^T \tilde{\Phi}(t, 0)^* \tilde{Z}(t)^* \tilde{Z}(t) \tilde{\Phi}(t, 0)\, dt\right)$$

is positive definite.

Proof. The proof is similar to the proof of Theorem 3.7. ▧

3.3. Duality. In general, it is difficult to find a differential system which synthesizes the dual system $\mathscr{A}^*$ (cf. Delfour and Mitter [11, Def. 12 and Thm. 13]). However, it is not too difficult to construct the dual system corresponding to the notions of "controllability at time T" and "observability at time T." The simultaneously controlled and observed dual of $\mathscr{A}$ is defined as follows:

$$\begin{aligned} &\frac{dx}{dt}(t) + A_{00}(t)^* x(t) + \sum_{i=1}^{N} \begin{cases} A_i(t - \theta_i)^* x(t - \theta_i), & t - \theta_i \leqq T \\ 0, & t - \theta_i > T \end{cases} \\ &+ \int_{-b}^{0} \begin{cases} A_{01}(t - \theta, \theta)^* x(t - \theta), & t - \theta \leqq T \\ 0, & t - \theta > T \end{cases} d\theta + Z(t)^* y(t) = 0 \\ &\qquad\qquad \text{a.e. in } [0, T], \\ &x(T) = x_T \in H \text{ (evolution equation)}, \\ &\chi(t; x_T, y) = B(t)^* \phi^*(t; x_T, y) \text{ (observation map)}, \end{aligned} \tag{3.15}$$

where $\phi^*(\cdot\,; x_T, y)$ is the unique solution of (3.15) in $AC^2(0, T; H)$.

PROPOSITION 3.17. *System* $\mathscr{A}$ *is controllable at time* T *(resp. observable in* $[0, T]$*) if and only if system* $\mathscr{A}^*$ *is observable in* $[0, T]$ *(resp. controllable at* 0*)*. ▧

It is extremely important to notice that "controllability at time T" is a dual notion of "observability in $[0, T]$." It would have been extremely unpleasant to have "strong observability" in lieu of "observability."

4. The optimal control problem with a quadratic cost.

4.1. Formulation of the problem. Consider the controlled system (2.1). We fix the final time $T \in]0, \infty[$ and consider the solution of (2.1) in the interval

$[0, T]$. We also consider f to be given. The solution in $[0, T]$ corresponding to $h \in M^2(-b, 0; H)$ and $v \in L^2(0, T; U)$ is denoted by $x(\cdot\,; h, v)$. We associate with v and h the cost function $J(v, h)$ given by

$$\begin{aligned} J(v, h) &= (x(T; h, v)|Fx(T; h, v)) \\ &\quad + \int_0^T [(x(s; h, v)|Q(s)x(s; h, v)) + (v(s)|N(s)v(s))]\, ds \\ &\quad + 2\int_0^T (v(s)|m(s))\, ds + 2\int_0^T (x(s; h, v)|g(s))\, ds, \end{aligned} \tag{4.1}$$

where $g \in L^2(0, T; H)$, $m \in L^2(0, T; U)$, $F \in \mathscr{L}(H)$, $Q \in L^\infty(0, T; \mathscr{L}(H))$, $N \in L^\infty(0, T; \mathscr{L}(U))$, F, $Q(s)$ and $N(s)$ are positive symmetric transformations and there exists a constant $c > 0$ such that $(y|N(s)y) \geqq c\|y\|_U^2$ for all s in $[0, T]$.

For each h we shall show that there exists a unique control u which minimizes the cost function $J(v, h)$ over all v in $L^2(0, T; U)$. The minimizing control u will be completely characterized in terms of the adjoint system. We shall also show that the control u can be synthesized using a linear feedback law and that the minimum of the cost function can be expressed in terms of the initial datum h.

4.2. Existence of the optimal control; Necessary and sufficient conditions for optimality. The existence and uniqueness of the optimal control u minimizing the cost $J(v, h)$ is a direct consequence of the hypotheses of §§ 2 and 4.1 and two theorems of Lions (cf. [26, Thm. 1.1, p. 4, and Thm. 1.2, p. 7]). In summary, given a continuous bilinear form $\tilde{\pi}$ defined in a Hilbert space $\mathscr{U}$ (with norm $\|\cdot\|$) satisfying the properties

$$\tilde{\pi}(v, w) = \tilde{\pi}(w, v) \quad \text{for all } w, v \in \mathscr{U}, \tag{4.2a}$$

$$\tilde{\pi}(v, v) \geqq c\|v\|^2 \quad \text{for all } v \in \mathscr{U}, \quad c > 0, \tag{4.2b}$$

and a continuous linear form $\tilde{L}$ also defined in $\mathscr{U}$, we define the cost

$$\tilde{J}(v) = \tilde{\pi}(v, v) - 2\tilde{L}(v) \tag{4.3}$$

which is to be minimized over the closed convex subset $\mathscr{U}_{ad}$ of $\mathscr{U}$. For such a cost there exists a unique u in $\mathscr{U}_{ad}$ minimizing $\tilde{J}(v)$ and this element can be uniquely characterized by

$$\tilde{\pi}(u, v - u) \geqq \tilde{L}(v - u) \quad \text{for all } v \in \mathscr{U}_{ad}. \tag{4.4}$$

For fixed f the cost function $J(v, h)$ given by (4.1) is of the form

$$J(v, h) = \pi(v, v) - 2L_h(v) + c(h), \tag{4.5}$$

where π and L_h satisfy the same hypotheses as $\tilde{\pi}$ and $\tilde{L}$ and $c(h)$ is a constant which solely depends on h. If $y(\cdot\,; w) = x(\cdot\,; 0, u + w) - x(\cdot\,; 0, u)$ is the solution of system $\mathscr{A}$ with $f = 0$, $h = 0$ and control w, a straightforward computation will show that inequality (4.4) becomes

$$\begin{aligned} &\int_0^T [(Q(s)x(s; h, u) + g(s)|y(s; w)) + (N(s)u(s) + m(s)|w(s))]\, ds \\ &\qquad + (Fx(T; h, u)|y(T; w)) \geqq 0 \quad \text{for all } w \in L^2(0, T; U). \end{aligned} \tag{4.6}$$

In order to improve the above characterization, we introduce the adjoint system corresponding to $x(\cdot\,; h, v)$:

$$
\begin{aligned}
&\frac{dp}{ds}(s; h, v) + A_{00}(s)^* p(s; h, v) + \sum_{i=1}^{N} \begin{cases} A_i(s-\theta_i)^* p(s-\theta_i; h, v), & s-\theta_i \leqq T \\ 0, & s-\theta_i > T \end{cases} \\
&\quad + \int_{-b}^{0} \begin{cases} A_{01}(s-\theta, \theta)^* p(s-\theta; h, v), & s-\theta \leqq T \\ 0, & s-\theta > T \end{cases} d\theta \\
&\quad + Q(s)x(s; h, v) + g(s) = 0 \quad \text{a.e. in } [0, T], \\
&p(T; h, v) = Fx(T; h, v).
\end{aligned} \tag{4.7}
$$

The notation $p(s; h, v)$ emphasizes the dependence of the adjoint solution on the control v and the initial datum h. From Lemma 3.3 in Delfour and Mitter [7] we know that by letting $y(s; w) = x(s; h, w) - x(s; h, 0)$,

$$
\begin{aligned}
&\mathscr{H}(T; T, (0 \circ y(\cdot\,; w))_T, p(\cdot\,; h, u)) - \mathscr{H}(0; T, (0 \circ y(\cdot\,; w))_0, p(\cdot\,; h, u)) \\
&= \int_0^T \Bigg(p(s; h, u) \Big| \frac{dy}{ds}(s; w) - A_{00}(s) y(s; w) - \sum_{i=1}^{N} \begin{cases} A_i(s) y(s+\theta_i; w), & s+\theta_i \geqq 0 \\ 0, & s+\theta_i < 0 \end{cases} \\
&\quad - \int_{-b}^{0} \begin{cases} A_{01}(s, \theta)(s+\theta; w), & s+\theta \geqq 0 \\ 0, & s+\theta < 0 \end{cases} d\theta \Bigg) \\
&\quad + \int_0^T \Bigg(\frac{dp}{ds}(s; h, u) + A_{00}(s)^* p(s; h, u) \\
&\quad + \sum_{i=1}^{N} \begin{cases} A_i(s-\theta_i)^* p(s-\theta_i; h, u), & s-\theta_i \leqq T \\ 0, & s-\theta_i > T \end{cases} \\
&\quad + \int_{-b}^{0} \begin{cases} A_{01}(s-\theta, \theta)^* p(s-\theta; h, u), & s-\theta \leqq T \\ 0, & s-\theta > T \end{cases} d\theta | y(s; w) \Bigg) ds.
\end{aligned} \tag{4.8}
$$

Computing $Q(s)x(s; h, u) + g(s)$ in (4.6) by using (4.7) and (4.8), we can rewrite inequality (4.6) explicitly as

$$
\int_0^T (B(s)^* p(s; h, u) + N(s)u(s) + m(s) | w(s))\, ds \geqq 0 \quad \text{for all } w \in L^2(0, T; U). \tag{4.9}
$$

Hence the optimal control u is uniquely characterized by

$$
u(s) = -N(s)^{-1}[B(s)^* p(s; h, u) + m(s)] \quad \text{a.e. in } [0, T].
$$

Thus we have established the following result.

THEOREM 4.1 (Necessary and sufficient conditions of optimality). *Given h, there exists a unique control u which minimizes $J(v, h)$ in $L^2(0, T; U)$. The optimal control u is completely characterized by the identity*

$$
u(s) = -N(s)^{-1}[B(s)^* p(s; h) + m(s)] \quad \textit{a.e. in } [0, T], \tag{4.10}
$$

where $(p(\cdot\,; h), x(\cdot\,; h))$ is the unique pair of maps in $AC^2(0, T; H)$ which satisfies

the following system of equations:

$$\begin{aligned}\frac{dx}{ds}(s;h) &= A_{00}(s)x(s;h) + \sum_{i=1}^{N} A_i(s)\begin{Bmatrix}x(s+\theta_i;h), & s+\theta_i \geqq 0\\ h(s+\theta_i), & s+\theta_i < 0\end{Bmatrix}\\ &\quad + \int_{-b}^{0} A_{01}(s,\theta)\begin{Bmatrix}x(s+\theta;h), & s+\theta \geqq 0\\ h(s+\theta), & s+\theta < 0\end{Bmatrix} d\theta\\ &\quad - B(s)N(s)^{-1}[B(s)^*p(s;h) + m(s)] + f(s)\end{aligned} \tag{4.11}$$

a.e. in $[0, T]$,

$$x(0;h) = h(0);$$

$$\begin{aligned}&\frac{dp}{ds}(s;h) + A_{00}(s)^*p(s;h) + \sum_{i=1}^{N}\begin{Bmatrix}A_i(s-\theta_i)^*p(s-\theta_i;h), & s-\theta_i \leqq T\\ 0, & s-\theta_i > T\end{Bmatrix}\\ &+ \int_{-b}^{0}\begin{Bmatrix}A_{01}(s-\theta,\theta)^*p(s-\theta;h), & s-\theta \leqq T\\ 0, & s-\theta > T\end{Bmatrix} d\theta + Q(s)x(s;h) + g(s) = 0\end{aligned} \tag{4.12}$$

a.e. in $[0, T]$,

$$p(T;h) = Fx(T;h).$$

Proof. The proof of this theorem is an immediate consequence of the existence and uniqueness of the pair $(x(\cdot\,:h,v), p(\cdot\,;h,v))$ as solutions of (2.1) and (4.7) and the characterization given by (4.10). ▨

COROLLARY 4.2. *Let* $m = 0$ *in* (4.11). *The two-point boundary value problem* (4.11)–(4.12) *has a unique solution* $(x(\cdot\,;h), p(\cdot\,;h))$ *in* $AC^2(0, T; H) \times AC^2(0, T; H)$. ▨

Remark. A different type of boundary value problem for HDE can be found in a paper by A. Halanay [17].

4.3. "Decoupling" of the equations of Theorem 4.1; the operators D and P. In this section we consider the initial datum h to be fixed. Let $f'(r) = f(r) - B(r)N(r)^{-1}m(r)$ and $R(r) = B(r)N(r)^{-1}B(r)^*$. We shall write $x(r)$, $p(r)$ and $J(v)$ in place of $x(r;h)$, $p(r;h)$ and $J(v,h)$ as in Theorem 4.1. In order to "decouple" the system of equations (4.11)–(4.12) we consider the problem of § 4.2 in the interval $[s, T]$, $s \in [0, T[$ instead of $[0, T]$. In this case the solution of (2.1) in the interval $[s, T]$ is denoted by $\phi(\cdot\,;s,v)$ and the cost is defined by

$$\begin{aligned}J_s(v) &= (\phi(T;s,v)|F\phi(T;s,v))\\ &\quad + \int_s^T [(\phi(r;s,v)|Q(r)\phi(r;s,v)) + (v(r)|N(r)v(r))]\,dr\\ &\quad + 2\int_s^T [(\phi(r;s,v)|g(r)) + (v(r)|m(r))]\,dr.\end{aligned} \tag{4.13}$$

Corresponding to $\phi(\cdot\,;s,v)$, the solution of

$$\begin{aligned}\frac{d\phi}{dr}(r;s,v) &= A_{00}(r)\phi(r;s,v) + \sum_{i=1}^{N} A_i(r)\begin{Bmatrix}\phi(r+\theta_i;s,v), & r+\theta_i \geqq s\\ h(r+\theta_i-s), & r+\theta_i < s\end{Bmatrix}\\ &\quad + \int_{-b}^{0} A_{01}(r,\theta)\begin{Bmatrix}\phi(r+\theta;s,v), & r+\theta \geqq s\\ h(r+\theta-s), & r+\theta < s\end{Bmatrix} d\theta\\ &\quad + B(r)v(r) + f(r) \quad \text{a.e. in } [s, T],\\ \phi(s;s,v) &= h(0),\end{aligned} \tag{4.14}$$

we introduce the adjoint solution $\psi(\cdot\,; s, v)$ as the solution of

$$\begin{aligned}
&\frac{d\psi}{dr}(r; s, v) + A_{00}(r)^*\psi(r; s, v) + \sum_{i=1}^{N} \begin{Bmatrix} A_i(r-\theta_i)^*\psi(r-\theta_i; s, v), & r-\theta_i \leqq T \\ 0, & r-\theta_i > T \end{Bmatrix} \\
&\quad + \int_{-b}^{0} \begin{Bmatrix} A_{01}(r-\theta, \theta)^*\psi(r-\theta; s, v), & r-\theta \leqq T \\ 0, & r-\theta > T \end{Bmatrix} d\theta \qquad (4.15) \\
&\quad + Q(r)\phi(r; s, v) + g(r) = 0 \quad \text{a.e. in } [s, T], \\
&\psi(T; s, v) = F\phi(T; s, v).
\end{aligned}$$

We now obtain the analogue of Theorem 4.1 for the optimal control $u \in L^2(s, T; U)$ which is characterized by

$$u(r) = -N(r)^{-1}[B(r)^*\psi(r; s) + m(r)] \quad \text{a.e. in } [s, T], \qquad (4.16)$$

where the pair $(\phi(\cdot\,; s),\ \psi(\cdot\,; s)) = (\phi(\cdot\,; s, u),\ \psi(\cdot\,; s, u))$ is the solution in $AC^2(s, T; H) \times AC^2(s, T; H)$ of the coupled system

$$\begin{aligned}
&\frac{d\phi}{dr}(r; s) = A_{00}(r)\phi(r; s) + \sum_{i=1}^{N} A_i(r) \begin{Bmatrix} \phi(r+\theta_i; s), & r+\theta_i \geqq s \\ h(r+\theta_i - s), & r+\theta_i < s \end{Bmatrix} \\
&\quad + \int_{-b}^{0} A_{01}(r, \theta) \begin{Bmatrix} \phi(r+\theta; s), & r+\theta \geqq s \\ h(r+\theta-s), & r+\theta < s \end{Bmatrix} d\theta \qquad (4.17) \\
&\quad - R(r)\psi(r; s) + f'(r) \quad \text{a.e. in } [s, T], \\
&\phi(s, s) = h(0); \\
&\frac{d\psi}{dr}(r; s) + A_{00}(r)^*\psi(r; s) + \sum_{i=1}^{N} \begin{Bmatrix} A_i(r-\theta_i)^*\psi(r-\theta_i; s), & r-\theta_i \leqq T \\ 0, & r-\theta_i > T \end{Bmatrix} \\
&\quad + \int_{-b}^{0} A_{01}(r-\theta, \theta)^* \begin{Bmatrix} \psi(r-\theta; s), & r-\theta \leqq T \\ 0, & r-\theta > T \end{Bmatrix} d\theta \qquad (4.18) \\
&\quad + Q(r)\phi(r; s) + g(r) = 0 \quad \text{a.e. in } [s, T], \\
&\psi(T; s) = F\phi(T; s).
\end{aligned}$$

LEMMA 4.3. *The map*

$$\begin{aligned}
&(h, f, g, m) \mapsto (\phi(\cdot\,; s), \psi(\cdot\,; s)) \qquad (4.19) \\
&: M^2(-b, 0; H) \times L^2(s, T; H) \times L^2(s, T; H) \times L^2(s, T; U) \\
&\quad \to AC^2(s, T; H) \times AC^2(s, T; H)
\end{aligned}$$

is linear and continuous.

Proof. The map (4.19) is clearly linear. To show it is continuous, choose an arbitrary sequence $\{(h_n, f_n, g_n, m_n)\}$ in $M^2 \times L^2 \times L^2 \times L^2$ which converges to (h, f, g, m). Let $(\phi_n(\cdot\,; s, v), \psi_n(\cdot\,; s, v))$ (resp. $(\phi(\cdot\,; s, v), \psi(\cdot\,; s, v))$) be the solution of the system (4.14)–(4.15) for some $v \in L^2(s, T; U)$ and the data (h_n, f_n, g_n, m_n) (resp. (h, f, g, m)). For fixed v,

$$(h_n, f_n, g_n, m_n) \to (h, f, g, m) \Rightarrow \phi_n(\cdot\,; s, v) \to \phi(\cdot\,; s, v) \quad \text{in } AC^2(s, T; H) \qquad (4.20)$$

by [7, Cor. 2.7]. Let u_n (resp. u) be the optimal control for the cost $J_s^n(v)$ (resp. $J_s(v)$) defined in terms of (h_n, f_n, g_n, m_n) (resp. (h, f, g, m)). We clearly obtain

$$J_s^n(u_n) = \inf_{v \in L^2} J_s^n(v) \leqq J_s^n(u) \tag{4.21}$$

and

$$J_s^n(u) \to J_s(u) \tag{4.22}$$

by (4.20) with $v = u$ and the very construction of J_s^n and J_s.

Hence

$$\limsup J_s^n(u_n) \leqq \limsup J_s^n(u) = J_s(u) = \inf_{v \in L^2} J_s(v). \tag{4.23}$$

Because of the hypothesis on $N(r)$, there exist $c_1 > 0$ and $c_2 > 0$ for which

$$J_s^n(u_n) \geqq c_1 |u_n|^2 - c_2 |u_n|. \tag{4.24}$$

From the last inequality and (4.23) it is necessary that

$$u_n \in \text{bounded subset of } L^2 \quad \text{as } (h_n, f_n, g_n, m_n) \to (h, f, g, m). \tag{4.25}$$

By weak compactness there is a subsequence u_μ such that

$$u_\mu \to w \quad \text{in } L^2 \text{ weakly.} \tag{4.26}$$

Thus

$$\phi_\mu(\cdot\,; s, u_\mu) \to \phi(\cdot\,; s, w) \quad \text{in } AC^2(s, T; H) \text{ weakly} \tag{4.27}$$

and finally (by convexity of $J_s(v)$ in v),

$$\liminf J_s^\mu(u_\mu) \geqq J_s(w). \tag{4.28}$$

If we combine inequalities (4.23) and (4.28), we necessarily have $w = u$. The results are summarized below:

$$\begin{aligned} &u_n \to u \quad \text{in } L^2(s, T; H) \text{ weakly,} \\ &J_s^n(u_n) \to J_s(u); \end{aligned} \tag{4.29}$$

$$\begin{aligned} &\phi_n(\cdot\,; s, u_n) \to \phi(\cdot\,; s, u) \quad \text{in } AC^2(s, T; H) \text{ weakly,} \\ &\psi_n(\cdot\,; s, u_n) \to \psi(\cdot\,; s, u) \quad \text{in } AC^2(s, T; H) \text{ weakly.} \end{aligned} \tag{4.30}$$

This shows the continuity on $M^2 \times L^2 \times L^2 \times L^2$ (in the strong topology) into $AC^2(s, T; H) \times AC^2(s; T; H)$ (in the weak topology) of the map (4.11). Since all the spaces in presence are Hilbert this is sufficient to prove the theorem. ▧

Remark. The above is essentially Lions' proof [26, Lemma 4.2, pp. 148–150].

COROLLARY 4.4. *The map*

$$(h, f, g, m) \mapsto \psi(r; s) : M^2 \times L^2 \times L^2 \times L^2 \to H \tag{4.31}$$

is linear and continuous for $r \in [s, T]$. *Hence it has the representation*

$$\psi(r; s) = D(r, s)h + F(r, s)f + G(r, s)g + M(r, s)m \tag{4.32}$$

for

$$D(r, s) \in \mathscr{L}(M^2(-b, 0; H), H), \qquad F(r, s) \in \mathscr{L}(L^2(s, T; H), H),$$
$$G(r, s) \in \mathscr{L}(L^2(s, T; U), H) \quad \text{and} \quad M(r, s) \in \mathscr{L}(L^2(s, T; K), H). \quad ▧$$

In the remainder of this section we assume that f, g and m are fixed. In this case we may write

$$\psi(r;s) = D(r,s)h + d(r,s) \tag{4.33}$$

instead of (4.32).

LEMMA 4.5. *Let* $(x,p) = (x(\cdot\,;h), p(\cdot\,;h))$ *be the solution of the system* (4.11)–(4.12). *Then*

$$p(t) = D(t,s)\tilde{x}(s;h) + d(t,s) \tag{4.34}$$

for all pairs $s \leqq t$ *in* $[0,T]$, *where* $D(t,s)$ *and* $d(t,s)$ *are defined by the following rules*:

(i) *We solve the system*

$$\begin{aligned}\frac{d\beta}{dr}(r) &= A_{00}(r)\beta(r) + \sum_{i=1}^{N} A_i(r)\begin{cases}\beta(r+\theta_i), & r+\theta_i \geqq s\\ h(r+\theta_i - s), & r+\theta_i < s\end{cases}\\ &\quad + \int_{-b}^{0} A_{01}(r,\theta)\begin{cases}\beta(r+\theta), & r+\theta \geqq s\\ h(r+\theta-s), & r+\theta < s\end{cases} d\theta - R(r)\gamma(r)\\ &\qquad \text{a.e. in } [s,T] \quad \text{and} \quad \beta(s) = h(0),\end{aligned} \tag{4.35}$$

$$\begin{aligned}&\frac{d\gamma}{dr}(r) + A_{00}(r)^*\gamma(r) + \sum_{i=1}^{N}\begin{cases}A_i(r-\theta_i)^*\gamma(r-\theta_i), & r-\theta_i \leqq T\\ 0, & r-\theta_i > T\end{cases}\\ &\quad + \int_{-b}^{0}\begin{cases}A_{01}(r-\theta,\theta)^*\gamma(r-\theta), & r-\theta \leqq T\\ 0, & r-\theta > T\end{cases} d\theta + Q(r)\beta(r) = 0\\ &\qquad \text{a.e. in } [s,T] \quad \text{and} \quad \gamma(T) = F\beta(T);\end{aligned} \tag{4.36}$$

then

$$D(t,s)h = \gamma(t), \quad t \in [s,T].$$

(ii) *We solve the system*

$$\begin{aligned}\frac{d\eta}{dr}(r) &= A_{00}(r)\eta(r) + \sum_{i=1}^{N} A_i(r)\begin{cases}\eta(r+\theta_i), & r+\theta_i \geqq 0\\ 0, & r+\theta_i < 0\end{cases}\\ &\quad + \int_{-b}^{0} A_{01}(r,\theta)\begin{cases}\eta(r+\theta), & r+\theta \geqq 0\\ 0, & r+\theta < 0\end{cases} d\theta - R(r)\xi(r) + f'(r)\\ &\qquad \text{a.e. in } [s,T] \quad \text{and} \quad \eta(s) = 0,\end{aligned} \tag{4.37}$$

$$\begin{aligned}&\frac{d\xi}{dr}(r) + A_{00}(r)^*\xi(r) + \sum_{i=0}^{N}\begin{cases}A_i(r-\theta_i)^*\xi(r-\theta_i), & r-\theta_i \leqq T\\ 0, & r-\theta_i > T\end{cases}\\ &\quad + \int_{-b}^{0}\begin{cases}A_{01}(r-\theta,\theta)^*\xi(r-\theta), & r-\theta \leqq T\\ 0, & r-\theta > T\end{cases} d\theta + Q(r)\eta(r) + g(r) = 0\\ &\qquad \text{a.e. in } [s,T] \quad \text{and} \quad \xi(T) = F\eta(T);\end{aligned} \tag{4.38}$$

then

$$d(t,s) = \xi(t), \quad t \in [s,T].$$

Proof. $D(t, s)$ and $d(t, s)$ are clearly obtained from the rules (i) and (ii) of the lemma: it suffices to decompose the map $h \to \psi(r)$ into its linear part and its constant part. We only need to establish identity (4.34). Consider the system (4.14)–(4.15) with initial datum $\tilde{x}(s; h)$ (see Definition 2.1) at time s, where x is the solution of the system (4.11)–(4.12) with initial datum h. The solution is denoted by (ϕ, ψ).

We also define

$$\bar{\phi} \text{ (resp. } \bar{\psi}) = \text{restriction of } x \text{ (resp. } p) \text{ to } [s, T],$$

where $\bar{\phi}$ and $\bar{\psi}$ are the solutions of the system (4.14)–(4.15) with initial datum $\tilde{x}(s; h)$. By uniqueness, $(\phi, \psi) = (\bar{\phi}, \bar{\psi})$ and $p(t) = \bar{\psi}(t) = D(t, s)\tilde{x}(s; h) + d(t, s)$. This proves the lemma. ▧

Remark. The above is essentially Lions' original proof (cf. J. L. Lions [26, Lemma 4.3]).

COROLLARY 4.6. *Given* $s \in [0, T[$ *and* $h \in M^2(-b, 0; H)$, *the maps* $t \mapsto D(t, s)h$ *and* $t \mapsto d(t, s)$ *are in* $AC^2(s, T; H)$. ▧

DEFINITION 4.7. For all $s \in [0, T[$ and $\eta \in I(-b, 0)$, $s - \eta \leqq T$, let

$$P(s, \eta) = \begin{cases} D(s - \eta, s), & \eta \in [s - T, 0] \cap I(-b, 0), \\ 0, & \text{otherwise.} \end{cases} \tag{4.39}$$

This defines the operator $P(s) \in \mathcal{L}(M^2(-b, 0; H))$ in the natural way $(P(s)h)(\eta) = P(s, \eta)h$. Similarly, let

$$r(s, \eta) = \begin{cases} d(s - \eta, s), & \eta \in [s - T, 0] \cap I(-b, 0), \\ 0, & \text{otherwise;} \end{cases} \tag{4.40}$$

and this defines $r(s) \in M^2(-b, 0; H)$, where $(r(s))(\eta) = r(s, \eta)$.

Remark. The conclusions of Lemma 4.5 can also be written in state form. Let $\tilde{x}, \tilde{p}, \tilde{\gamma}$ and $\tilde{\xi}$ denote the state variables associated with the variables x, p, γ and ξ, respectively. We can write $\tilde{p}(s) = P(s)\tilde{x}(s) + r(s)$, where $P(s)h = \tilde{\gamma}(s)$ and $r(s) = \tilde{\xi}(s)$. Here $P(s)$ and $r(s)$ are defined directly.

4.4. The operator $\Pi(s)$ and the optimal cost; relations between Π and P. In this section we introduce the operator $\Pi(s)$ which characterizes the optimal cost. It is constructed from $D(s - \eta, s)$, $\eta \in I(-b, 0)$, $s - \eta \leqq T$, or simply from $P(s, \eta)$ (Definition 4.5). Since there is an isometric isomorphism κ between $M^2(-b, 0; H)$ and $H \times L^2(-b, 0; H)$ the operator $P(s, \eta)$ can be decomposed in the following way:

$$P(s, \eta)h = P^0(s, \eta)h^0 + P^1(s, \eta)h^1, \tag{4.41}$$

where $h \in M^2(-b, 0; H)$, $\kappa(h) = (h^0, h^1)$ (see [6, Prop. 2.1]), $P^0(s, \eta) \in \mathcal{L}(H)$ and $P^1(s, \eta) \in \mathcal{L}(L^2(-b, 0; H), H)$.

PROPOSITION 4.8. *Let* $f = g = 0$ *and* $m = 0$ *in system* (4.17)–(4.18). *We denote by* $(\phi(\cdot\,; s), \psi(\cdot\,; s))$ *(resp.* $(\bar{\phi}(\cdot\,; s), \bar{\psi}(\cdot\,; s))$*) its solution for the initial datum* h *(resp.* $\bar{h}$*). Then*

$$\begin{aligned} \mathcal{H}(s; T, h, D(\cdot, s)\bar{h}) = (\phi(T; s) | F\bar{\phi}(T; s)) + \int_s^T [(\bar{\psi}(r; s) | R(r)\psi(r; s)) \\ + (Q(r)\bar{\phi}(r; s) | \phi(r; s))]\, dr. \end{aligned} \tag{4.42}$$

The map

$$(4.43)\quad (h,\bar{h}) \mapsto \mathscr{H}(s;T,h,D(\cdot,s)\bar{h}) : M^2(-b,0;H) \times M^2(-b,0;H) \to R$$

is a continuous bilinear form which is symmetric and positive.

Proof. From Lemma 3.3 of [7] and equations (4.17)–(4.18),

$$(4.44)\quad \begin{aligned} &\mathscr{H}(T;T,(h\circ\phi(\cdot,s))_T,\bar{\psi}(\cdot\,;s)) - \mathscr{H}(s;T,h,\bar{\psi}(\cdot\,;s)) \\ &\qquad = -\int_s^T [(\bar{\psi}(r;s)|R(r)\psi(r;s)) + (Q(r)\bar{\phi}(r;s)|\phi(r;s))]\,dr, \end{aligned}$$

where $\mathscr{H}$ is given by Definition 3.2 in [7]. But

$$(4.45)\quad \mathscr{H}(T;T,(h\circ\phi(\cdot\,;s))_T,\bar{\psi}(\cdot\,;s)) = (\phi(T;s)|\bar{\psi}(T;s)) = (\phi(T;s)|F\bar{\phi}(T;s))$$

and we obtain (4.42) from identity (4.44). The map (4.43) is clearly bilinear and continuous since the map $h \mapsto (\phi(\cdot\,;s),\psi(\cdot\,;s))$ (resp. $\bar{h} \mapsto (\bar{\phi}(\cdot\,;s),\bar{\psi}(\cdot\,;s))$ is linear and continuous by Lemma 4.1. The symmetry and positivity of the map (4.43) follow from the symmetry and positivity of the operators F, $R(r)$ and $Q(r)$. ▧

COROLLARY 4.9. (i) *The map* (4.43) *can be written in a unique way in terms of the transformation* $\Pi(s) \in \mathscr{L}(M^2(-b,0;H))$:

$$(4.46)\quad \mathscr{H}(s;T,h,D(\cdot,s)\bar{h}) = (\Pi(s)\bar{h}|h)_{M^2}.$$

(ii) $\Pi(s)$ *is equivalent to the matrix of operators*

$$(4.47)\quad \begin{pmatrix} \Pi^{00}(s) & \Pi^{01}(s) \\ \Pi^{10}(s) & \Pi^{11}(s) \end{pmatrix},$$

where $\Pi^{00}(s) \in \mathscr{L}(H)$, $\Pi^{01}(s) \in \mathscr{L}(L^2,H)$, $\Pi^{11}(s) \in \mathscr{L}(L^2)$ *and* $\Pi^{10}(s) \in \mathscr{L}(H,L^2)$.

(iii) *Moreover,*

$$(4.48)\quad \Pi^{00}(s)^* = \Pi^{00}(s) \geqq 0,$$

$$(4.49)\quad \Pi^{01}(s) = \Pi^{10}(s)^*,$$

$$(4.50)\quad \Pi^{11}(s)^* = \Pi^{11}(s) \geqq 0.$$

(iv) *Let* u *be the optimal control obtained under the hypotheses of Proposition* 4.6. *The optimal cost can be written as*

$$(4.51)\quad J_s^h(u) = (\Pi^{00}(s)h^0|h^0) + 2(\Pi^{01}(s)h^1|h^0) + (\Pi^{11}(s)h^1|h^1)_{L^2}.$$

Moreover, there exists a constant $c > 0$ *such that*

$$(4.52)\quad \|\Pi(s)h\|_{M^2} \leqq c\|h\|_{M^2}.$$

Proof. (i)–(iv) are obvious. The inequality in (iv) follows from the positivity and symmetry of the operator $\Pi(s)$ and

$$|(\Pi(s)h|h)_{M^2}| = J_s^h(u) \leqq J_s^h(0) \leqq c\|h\|_{M^2}^2. \quad ▧$$

The operator $\Pi(s)$ can now be expressed in terms of $P(s,\eta)$, $\eta \in I(-b,0)$, $s - \eta \leqq T$. In doing this we obtain further information on $\Pi(s)$.

COROLLARY 4.10. (i) $\Pi^{00}(s) = P^0(s,0), \Pi^{01}(s) = P^1(s,0)$ *and* $(\Pi^{10}(s)h^0)(\alpha) = \Pi^{10}(s,\alpha)h^0$, *where*

$$\Pi^{01}(s,\alpha) = \sum_{i=1}^{N} \begin{cases} A_i(s+\alpha-\theta_i)*P^0(s,\theta_i-\alpha), & \alpha+s-T<\theta_i\leqq\alpha \\ 0, & \text{otherwise} \end{cases} + \int_{\max\{-b,\alpha+s-T\}}^{\alpha} A_{01}(s+\alpha-\theta,\theta)*P^0(s,\theta-\alpha)\,d\theta. \tag{4.53}$$

(ii) *The kernel* $\Pi^{01}(s,\alpha)$ *of* $\Pi^{01}(s)$ *is equal to* $\Pi^{10}(s,\alpha)$*.

(iii)

$$(\Pi^{11}(s)h^1)(\alpha) = \sum_{i=1}^{N} \begin{cases} A_i(s+\alpha-\theta_i)*P^1(s,\theta_i-\alpha)h^1, & \alpha+s-T<\theta_i\leqq\alpha \\ 0, & \text{otherwise} \end{cases} + \int_{\max\{-b,\alpha+s-T\}}^{\alpha} A_{01}(s+\alpha-\theta,\theta)*P^1(s,\theta-\alpha)h^1\,d\theta. \tag{4.54}$$

If we now go back to the system (2.1) *defined in* $[0,T]$, *the minimizing control* $u(s)$ *at time* s *is given by*

$$u(s) = -N(s)^{-1}\left[B(s)^*\left[\Pi^{00}(s)x(s) + \int_{-b}^{0}\Pi^{10}(s,\alpha)^*\tilde{x}(s;h,u)(\alpha)\,d\alpha + r(s,0)\right] + m(s)\right]. \tag{4.55}$$

∎

5. Operational differential equation of Riccati type for the operators $P(t)$ and $\Pi(t)$. So far we have established the existence of operators $P(t)$ and $\Pi(t)$ in $\mathscr{L}(M^2)$ and studied their properties. We have also shown how $P(t)$ and $\Pi(t)$ can be indirectly computed. In this section we show that $P(t)$ and $\Pi(t)$ satisfy operational differential equations of Riccati type. In order to study $P(t)$ and $\Pi(t)$ we assume that in (2.1), $f = 0$, and in (4.1) that $m = 0$ and $g = 0$ (Proposition 4.6).

5.1. Formal derivation of an operational differential equation for $\Pi(t)$. We use the fact that there is an isometric isomorphism κ between $M^2(-b,0;H)$ and $H \times L^2(-b,0;H)$, where $\kappa(h) = (h^0,h^1)$. Since $\Pi(s) \in \mathscr{L}(M^2(-b,0;H))$, we use the above isomorphism and write

$$\begin{aligned} [\Pi(s)h]^0 &= \Pi^0(s)h, \quad \Pi^0(s) \in \mathscr{L}(M^2,H), \\ [\Pi(s)h]^1 &= \Pi^1(s)h, \quad \Pi^1(s) \in \mathscr{L}(M^2,L^2). \end{aligned} \tag{5.1}$$

We denote by (p,x) (resp. $(\bar{p},\bar{x})$) the solution of system (4.11)–(4.12) with initial datum h (resp. $\bar{h}$). We use the notation (see Definition 2.1)

$$x_s = \tilde{x}(s;h) \quad (\text{resp. } \bar{x}_s = \tilde{\bar{x}}(s;h)). \tag{5.2}$$

From (4.34), Definition 4.7, Corollary 4.10 and (5.1),

$$p(s) = \Pi^0(s)x_s \quad (\text{resp. } \bar{p}_s = \Pi^0(s)\bar{x}_s). \tag{5.3}$$

Define operators $\tilde{R}(t)$, $\tilde{Q}(t)$ and $\tilde{F}$ in $\mathscr{L}(M^2)$ as

$$[\tilde{R}(t)h](\alpha) = \begin{cases} R(t)h(0), & \alpha = 0, \\ 0, & \text{otherwise}; \end{cases} \tag{5.4}$$

$$[\tilde{Q}(t)h](\alpha) = \begin{cases} Q(t)h(0), & \alpha = 0, \\ 0, & \text{otherwise}; \end{cases} \tag{5.5}$$

$$[\tilde{F}h](\alpha) = \begin{cases} Fh(0), & \alpha = 0, \\ 0, & \text{otherwise}. \end{cases} \tag{5.6}$$

It is easy to verify the following:

$$[\tilde{R}(t)\Pi(t)x_t](0) = R(t)\Pi^0(t)x_t = R(t)p(t), \tag{5.7}$$

and

$$(p(t)|R(t)\bar{p}(t))_H = (\Pi(t)x_t|\tilde{R}(t)\Pi(t)\bar{x}_t), \tag{5.8}$$

$$(x(T)|F\bar{x}(T))_H = (x_T|\tilde{F}\bar{x}_T)_{M^2}, \tag{5.9}$$

$$(x(t)|Q(t)\bar{x}(t))_H = (x_t|\tilde{Q}(t)\bar{x}_t)_{M^2}. \tag{5.10}$$

Then from (2.16) in Theorem 2.3, (4.34) and (5.4),

$$\begin{aligned} &\frac{dx_t}{dt} = \tilde{A}(t)x_t - \tilde{R}(t)\Pi(t)x_t, \quad t \in [0, T], \\ &x_0 = h. \end{aligned} \tag{5.11}$$

Equations (5.4) through (5.11) can also be written with $\bar{h}$ and $\bar{x}$ in place of h and x. From (4.42), (4.46) and (5.8),

$$(x_s|\Pi(s)\bar{x}_s) = (x_T|\tilde{F}\bar{x}_T) + \int_s^T [(x_r|\tilde{Q}(r)\bar{x}_r) + (\Pi(r)x_r|\tilde{R}(r)\Pi(r)\bar{x}_r)]\, dr \tag{5.12}$$

and

$$(x_T|\Pi(T)\bar{x}_T) = (x_T|\tilde{F}\bar{x}_T). \tag{5.13}$$

Formal differentiation of both sides of (5.12) and use of (5.11) yields

$$([\dot{\Pi}(s) + \Pi(s)\tilde{A}(s) + \tilde{A}(s)^*\Pi(s) - \Pi(s)\tilde{R}(s)\Pi(s) + \tilde{Q}(s)]x_s|\bar{x}_s)_{M^2} = 0.$$

Since this has to be true for all x_s and $\bar{x}_s$, we get

$$\begin{aligned} &\dot{\Pi}(s) + \Pi(s)\tilde{A}(s) + \tilde{A}(s)^*\Pi(s) - \Pi(s)\tilde{R}(s)\Pi(s) + \tilde{Q}(s) = 0, \\ &\Pi(T) = \tilde{F}, \end{aligned} \tag{5.14}$$

where $\tilde{A}(s)^*$ is the M^2-adjoint of $\tilde{A}(s)$.

5.2. Interpretation of equation (5.14). The first question is to determine in what sense equation (5.14) has a solution. There are two ways to proceed: either to study (5.14) directly or to study an equivalent integral equation. In the first situation we can apply certain results of Da Prato [4], but we need further assumptions.

In the second situation we study an equivalent integral equation rather than the differential equation directly.

5.3. Direct study of equation (5.14). In order to apply Da Prato's results to equation (5.14) we need further hypotheses:

(i) A_{00}, A_i, A_{01} and B in (2.1) and Q and R in (4.1) do not depend on t;
(ii) there exist $\omega \geqq 0$ and $K > 0$ such that

$$\|\tilde{\Phi}(t)\|_{\mathscr{L}(M^2)} \leqq K e^{-\omega t} \quad \text{for all } t \geqq 0;$$

(iii) $\tilde{F} = 0$.

5.3.1. Results of Da Prato. We now state the results of Da Prato (cf. [4, Thms. 7.5 and 7.6]) which are of interest to us.

Let X be a Banach space. Let $\mathscr{L}(X, X)$ denote the algebra of bounded linear operators on X, $\mathscr{L}_s(X, X)$ the space $\mathscr{L}(X, X)$ endowed with the topology of simple convergence in X. Let M and N be two unbounded operators in X which are infinitesimal generators of strongly continuous semigroups $t \to e^{tM}$ and $t \to e^{tN}$ respectively. Moreover, we assume that there exist positive constants K_M, K_N and $\omega_N \in R_+$ such that

$$\|e^{tM}\| \leqq K_M \|e^{tN}\| \leqq K_N e^{-\omega_N t} \quad \text{for all } t \in R_+.$$

We consider the equation

$$\begin{aligned} &\frac{dU}{dt} - MU(t) - U(t)N - f(U(t)) = V(t), \\ &U(0) = 0, \quad t \in [0, T], \quad T \in R_+. \end{aligned} \tag{5.15}$$

In the above, f is holomorphic in an open set Ω of the complex plane containing the origin and $V \in C(0, T; \mathscr{L}_s(X, X))$.

DEFINITION 5.1. $U \in C(0, T; \mathscr{L}_s(X, X))$ is said to be a *weak solution of* (5.15) if there exists a sequence $\{U_n\}$ in $C^1(0, T; \mathscr{L}_s(X, X))$ such that[1]

(i) $U_n(0) = 0$ for all $n \in N$;
(ii) $U_n(t)x \in \mathscr{D}(M)$ (domain of M) for all $x \in X$, $t \in R_+$ and $t \mapsto MU_n(t) \in C(0, T; \mathscr{L}_s(X, X))$;
(iii) $U_n(t)N$ can be extended to a bounded operator $\overline{U_n(t)N}$, $t \in R_+$ and $t \mapsto \overline{U_n(t)N} \in C(0, T; \mathscr{L}_s(X, X))$;
(iv) $U_n \to U$ and $dU_n/dt - MU_n - \overline{U_nN} - f(U_n) \to V$ in $C(0, T; \mathscr{L}_s(X, X))$.

THEOREM 5.2 (Da Prato). *Let* $T_0 \in R_+$, $V \in C(0, T_0; \mathscr{L}_s(X, X))$. *Then there exists a* $T \leqq T_0$ *such that* (5.15) *has a unique weak solution in* $[0, T]$. ▨

Suppose further that there exists a constant $C \in R_+$ such that if $T \in [0, T_0]$ and U is a solution of (5.15) in $[0, T]$, we have

$$\|U(t)\| \leqq C \quad \text{for all } t \in [0, T].$$

THEOREM 5.3 (Da Prato). *Equation* (5.15) *has a unique global weak solution.* ▨

[1] $C^1(0, T; \mathscr{L}_s(X, X))$ = space of functions with values in $\mathscr{L}_s(X, X)$ which are strongly differentiable.

5.3.2. Existence of a global solution for (5.14). In view of the fact that $\tilde{A}$ and $\tilde{A}^*$ are infinitesimal generators of strongly continuous semigroups which are adjoint to each other and $f(\Pi) = \Pi\tilde{R}\Pi$, $\tilde{R}$ being a bounded positive operator, all the assumptions of Theorem 5.2 are satisfied and we can conclude that when $F = 0$ there exists a unique weak solution of (5.15) locally. Finally, in view of the a priori bound (4.52) we can conclude that the local solution is also global.

5.4. Study of equation (5.14) via an equivalent integral equation. We now derive an integral equation equivalent to (5.14).

PROPOSITION 5.4. *The operator* Π *defined in Corollary* 4.9 *is the unique minimal*[2] *solution of the following system*:

$$\begin{aligned} &\frac{\partial}{\partial r}\tilde{\Phi}(r,s) = [\tilde{A}(r) - \tilde{R}(r)\Pi(r)]\tilde{\Phi}(r,s) \quad \text{a.e. in } [s,T], \\ &\tilde{\Phi}(s,s) = I, \end{aligned} \tag{5.16}$$

$$\Pi(s) = \tilde{\Phi}(T,s)^*\tilde{F}\tilde{\Phi}(T,s) + \int_s^T \tilde{\Phi}(r,s)^*[\tilde{Q}(r) + \Pi(r)\tilde{R}(r)\Pi(r)]\tilde{\Phi}(r,s)\,dr. \tag{5.17}$$

Proof. (i) We start with Proposition 4.8. We know that for $f = g = 0$ and $m = 0$ system (4.17)–(4.18) has a unique solution and that

$$\psi(r,s) = P^0(r)\tilde{\phi}(r,s) = \Pi^0(r)\tilde{\phi}(r,s)$$

(cf. Lemma 4.5 and Definition 4.7).

Equation (4.17) can now be written in state form:

$$\frac{\partial}{\partial r}\tilde{\phi}(r,s) = [\hat{A}(r) - \tilde{R}(r)\Pi(r)]\tilde{\phi}(r,s) \quad \text{in } [s,T],$$

$$\tilde{\phi}(s,s) = h.$$

This clearly generates the semigroup $\tilde{\Phi}(r,s)$. Now using (4.42) in Proposition 4.8 and (4.46) in Corollary 4.9, we obtain

$$\begin{aligned} (h|\Pi(s)\bar{h}) = {} & (\tilde{\Phi}(T,s)h|\tilde{F}\tilde{\Phi}(T,s)\bar{h}) \\ & + \int_s^T [(\Pi(r)\tilde{\Phi}(r,s)h|\tilde{R}(r)\Pi(r)\tilde{\Phi}(r,s)\bar{h}) \\ & + (\tilde{\Phi}(r,s)h|\tilde{Q}(r)\tilde{\Phi}(r,s)\bar{h})]\,dr. \end{aligned}$$

This is sufficient to show that Π is a solution of system (5.16)–(5.17).

(ii) Consider another solution $\tilde{\Pi}$ of (5.16)–(5.17). Fix a time $s \in [0, T[$. This corresponds to the feedback control

$$\tilde{u}(t) = -N^{-1}B^*\tilde{\Pi}^0(t), \quad t \in [s,T],$$

in the time interval $[s,T]$. However, by definition,

$$(\Pi(s)h|h) = \min_{v \in L^2(s,T;U)} J_s(v) \leqq J_s(\tilde{u}) = (\tilde{\Pi}(s)h|h).$$

This proves the minimality property. ▨

[2] That is, the corresponding control gives us the minimal cost.

COROLLARY 5.5. *For all h and $\bar{h}$ in $\mathscr{D}$ the operator Π defined in Corollary* 4.9 *is a positive self-adjoint solution of the following system*:

$$\begin{aligned}&\frac{d}{dt}(\bar{h}|\Pi(t)h) + (\tilde{A}(t)\bar{h}|\Pi(t)h) + (\bar{h}|[\Pi(t)\tilde{A}(t) - \Pi(t)\tilde{R}(t)\Pi(t)\\ &\qquad + \tilde{Q}(t)]h) = 0 \quad \textit{a.e. in } [0, T],\\ &\Pi(T) = \tilde{F}.\end{aligned} \tag{5.18}$$

Proof. Let $\tilde{\tilde{\phi}}(t, s) = \tilde{\Phi}(t, s)\bar{h}$ and $\tilde{\phi}(t, s) = \tilde{\Phi}(t, s)h$. Using the fact that $\tilde{\phi}(r, s) = \tilde{\Phi}(r, t)\tilde{\phi}(t, s)$, we compute $(\tilde{\tilde{\phi}}(t, s)|\Pi(t)\tilde{\phi}(t, s))$ from (5.17):

$$\begin{aligned}(\tilde{\tilde{\phi}}(t, s)|\Pi(t)\tilde{\phi}(t, s)) = (\tilde{\tilde{\phi}}(T, s)|\tilde{F}\tilde{\phi}(T, s)) + \int_t^T (\tilde{\tilde{\phi}}(r, s)|[\tilde{Q}(r)\\ + \Pi(r)\tilde{R}(r)\Pi(r)]\tilde{\phi}(r, s))\, dr.\end{aligned}$$

We differentiate the above expression with respect to t and set s equal to t in the resulting expression to obtain (5.18). ▨

5.5. Integral and differential equations for operator $P(t)$. In this section we derive equations for operator $P(t)$. We shall use the operators $D(t, s)$, $P(t, \alpha)$ and $P(t)$ of Definition 4.7. Given $G \in \mathscr{L}(M^2)$ we denote by $G^0 \in \mathscr{L}(M^2, H)$ the operator defined by $G^0h = (Gh)(0)$ and by

$$\begin{bmatrix} G^{00} & G^{01} \\ G^{10} & G^{11} \end{bmatrix}$$

the corresponding matrix of operators defined on $H \times L^2(-b, 0; H)$.

PROPOSITION 5.6. *The operator P of Definition* 4.7 *is the unique solution of the following system of equations*:

$$\begin{aligned}&P^0(t) = \Psi(T, t)\tilde{F}^0\tilde{\Phi}(T, t) + \int_t^T \Psi(r, t)[\tilde{Q}(r)^0 + P^{00}(r)R(r)P^0(r)]\tilde{\Phi}(r, t)\, dr,\\ &[P(t)h](\alpha) = \begin{cases} P^0(t - \alpha)\tilde{\Phi}(t - \alpha, t)h, & t - \alpha \leqq T,\\ 0, & \textit{otherwise}.\end{cases}\end{aligned} \tag{5.19}$$

$\tilde{\Phi}(t, s)$ is the semigroup generated by the solutions of

$$\begin{aligned}&\frac{dz}{dt}(t) = [\tilde{A}(t) - \tilde{R}(t)P(t)]z(t) \quad \textit{a.e. in } [s, T],\\ &z(s) = h \in \mathscr{D}, \quad s \in [0, T[,\\ &\tilde{\Phi}(t, s)h = z(s),\end{aligned} \tag{5.20}$$

and $\Psi(t, s) \in \mathscr{L}(H)$ is generated by the solutions of

$$\begin{aligned}&\frac{dy}{ds}(s) + [\tilde{A}_t^0(s) - P^0(s)\tilde{R}(s)]y_s = 0 \quad \textit{a.e. in } [0, t], \quad y(t) = k^0,\\ &\Psi(t, s)k^0 = y(s),\end{aligned} \tag{5.21}$$

where $y_s(\theta)$ is equal to $y(s - \theta)$ when $s - \theta \leqq t$ and 0 when $s - \theta > t$.

Proof. (i) Notice that (5.20) and (5.21) are only functions of $P^0(t)$. Assume that $P^0(t)$ of Definition 4.7 is the unique solution of the first equation of (5.19). We have shown that

$$\psi(r,s) = D(r,s)h = D(r,r)\tilde{\phi}(r,s), \quad r \geqq s$$

(Corollary 4.4 and Lemma 4.5). The above equation can be rewritten

$$D(r,s)h = D(r,r)\Phi(r,s)h.$$

We let $r = t - \alpha$ and $s = t$ in the above equation and use Definition 4.7 to obtain the second equation of (5.19).

(ii) *Uniqueness.* Let $\bar{P}^0(t)$ be a solution of systems (5.19)–(5.21). Let $\tilde{\phi}(\cdot, s)$ be the solution in $[s, T]$ of the following equation:

$$\begin{aligned} &\frac{\partial}{\partial t}\tilde{\phi}(t,s) = [\tilde{A}(t) - \tilde{R}(t)\bar{P}(t)]\tilde{\phi}(t,s) \quad \text{in } [s,T], \\ &\tilde{\phi}(s,s) = h. \end{aligned} \tag{5.22}$$

By definition,

$$\tilde{\phi}(t,s) = \tilde{\Phi}(t,s)h. \tag{5.23}$$

We define

$$\psi(t,s) = \bar{P}^0(t)\tilde{\phi}(t,s). \tag{5.24}$$

We use (5.21) to obtain

$$\psi(t,s) = \Psi(T,t)F\phi(T,s) + \int_t^T \Psi(r,t)[\tilde{Q}(r)^0 + \bar{P}^{00}(r)R(r)\bar{P}^0(r)]\tilde{\phi}(r,s)\,dr. \tag{5.25}$$

We differentiate with respect to t both sides of (5.25):

$$\begin{aligned} &\frac{\partial}{\partial t}\psi(t,s) + \tilde{A}_T^0(t)\tilde{\psi}(t,s) + Q(t)\phi(t,s) = 0 \quad \text{in } [s,T], \\ &\psi(T,s) = F\phi(T,s), \end{aligned} \tag{5.26}$$

where $\tilde{\psi}(t,s)(\theta)$ is equal to $\psi(t-\theta, s)$ when $t - \theta \leqq T$ and 0 when $t - \theta > T$.

We can also rewrite (5.22) using (5.24):

$$\begin{aligned} &\frac{\partial}{\partial t}\tilde{\phi}(t,s) = \tilde{A}(t)\tilde{\phi}(t,s) - \tilde{R}(t)\tilde{\psi}(t,s) \quad \text{in } [s,T], \\ &\tilde{\phi}(s,s) = h. \end{aligned} \tag{5.27}$$

System (5.26)–(5.27) is the optimality system (4.17)–(4.18) and we know (Lemma 4.5 and Definition 4.7) that

$$\psi(s,s) = P^0(s)h. \tag{5.28}$$

Let $t = s$ in (5.24) and (5.28):

$$\bar{P}^0(s)h = \psi(s,s) = P^0(s)h. \tag{5.29}$$

Since this is true for all $s \in [0, T[$, we have established that a solution (if it exists) of system (5.19)–(5.21) is necessarily unique and equal to P^0.

(iii) *Existence.* Let P be as in Definition 4.7. The optimality system (4.17)–(4.18) can be put in the form (5.26)–(5.27) and (5.28) is true by Lemma 4.5 and Definition 4.7. As a result, $\tilde{\Phi}(t,s)$ and $\Psi(t,s)$ are well-defined and equation (5.26)

can be rewritten as follows:

$$\frac{\partial}{\partial t}\psi(t,s) + [\tilde{A}_T^0(t) - P^0(t)\tilde{R}(t)]\tilde{\psi}(t,s) + [\tilde{Q}(t)^0 + P^0(t)\tilde{R}(t)P(t)]\tilde{\phi}(t,s) = 0 \quad \text{a.e. in } [s,T], \tag{5.30}$$

$$\psi(T,s) = F\phi(T,s). \tag{5.31}$$

Using (2.22) we can rewrite system (5.30)–(5.31) in integral form:

$$\psi(t,s) = \Psi(T,t)F\phi(T,s) + \int_t^T \Psi(r,t)(\tilde{Q}(r)^0 + P^0(r)\tilde{R}(r)P(r))\tilde{\phi}(r,s)\,dr. \tag{5.32}$$

By using the relations

$$\tilde{\phi}(t,s) = \tilde{\Phi}(t,s)h \quad \text{and} \quad P^0(t)\tilde{\phi}(t,s) = \psi(t,s), \tag{5.33}$$

(5.32) becomes

$$P^0(t)\tilde{\phi}(t,s) = \left[\Psi(T,t)F\Phi(t,s) + \int_t^T \Psi(r,t)(\tilde{Q}(r)^0 + P^0(r)\tilde{R}(r)P(r))\tilde{\Phi}(r,s)\,dr\right]h. \tag{5.34}$$

Let $t = s$ in (5.34) and use the fact that $\tilde{\phi}(s,s) = h$ to obtain (5.21). This shows that P^0 is a solution of system (5.19)–(5.21). ▧

COROLLARY 5.7. *The operators P^0 and D of Definition* 4.7 *are solutions of the following coupled system*:

$$D(r,s) = P^0(r)\tilde{\Phi}(r,s), \quad r \geqq s, \tag{5.35}$$

and for all h,

$$\begin{aligned}\frac{d}{dt}[P^0(t)h] &+ [P^0(t)\tilde{A}(t) + A_{00}(t)*P^0(t) \\ &+ \sum_{i=1}^{N}\begin{cases}A_i(t-\theta_i)*D(t-\theta_i,t), & t-\theta_i \leqq T\\ 0, & \text{otherwise}\end{cases} \\ &+ \int_{-b}^{0}\begin{cases}A_{01}(t-\theta,\theta)*D(t-\theta,t), & t-\theta \leqq T\\ 0, & \text{otherwise}\end{cases}d\theta - P^{00}(t)R(t)P^0(t) \\ &+ \tilde{Q}(t)^0]h = 0 \quad \textit{a.e. in } [0,T],\end{aligned} \tag{5.36}$$

$$P^0(T) = \tilde{F}^0.$$

Proof. Consider equation (5.30). Using (5.28) and (5.20) we can compute

$$\frac{\partial\psi}{\partial t}(t,s) = \frac{\partial}{\partial u}[P^0(u)\tilde{\phi}(t,s)]_{u=t} + P^0(t)[\tilde{A}(t) - \tilde{R}(t)P(t)]\tilde{\phi}(t,s). \tag{5.37}$$

Let $s = t$ in (5.30) and (5.37). Using the definition of $\tilde{A}_T^0$ and $\tilde{\psi}(t,t)$ we obtain (5.36). ▧

Remark. $P(t)$ can be obtained from $D(r,s)$ (Definition 4.7):

$$[P(t)h](\alpha) = P(t,\alpha)h = D(t-\alpha,t)h, \quad t-\alpha \leqq T.$$

To complete the picture we need an equation for D. This equation can be formally obtained provided that the semigroup $\tilde{\Phi}(r, s)$ of system (5.20) is a solution of the following equation:

$$\frac{\partial}{\partial s}\tilde{\Phi}(r,s)^* + [\tilde{A}(s) - \tilde{R}(s)P(s)]^*\tilde{\Phi}(r,s)^* = 0 \quad \text{a.e. in } [0,r],$$

$$\tilde{\Phi}(r,r)^* = I.$$

This is equivalent to postulating the existence of a topological adjoint system for system (5.20). Under this hypothesis we formally differentiate (5.35) with respect to s to obtain the desired differential equation for D:

$$\frac{\partial}{\partial s}D(r,s) + D(r,s)[\tilde{A}(s) - \tilde{R}(s)P(s)] = 0 \quad \text{a.e. in } [0,r],$$

$$D(r,r) = P^0(r).$$

Let $(r, s) = (t - \alpha, t)$ in the above equation. Since $P(t, \alpha) = D(t - \alpha, t)$ we can obtain the following differential equation for $P(t, \alpha)$:

$$\left[\frac{\partial}{\partial t} + \frac{\partial}{\partial \alpha}\right]P(t,\alpha) + P(t,\alpha)[\tilde{A}(t) - \tilde{R}(t)P(t)] = 0$$

in the region $\{(t, \alpha) \in [0, T] \times I(-b, 0) | t - \alpha \leqq T\}$ with boundary conditions

$$P(s,0) = P^0(s), \quad s \in [0,T].$$

For completeness we also include the following result which is obtained by decomposition of (5.17) and the first equation of (5.19).

COROLLARY 5.8. *The operator* Π *defined in Corollary* 4.9 *is the unique solution of the following system of equations*:

$$\begin{aligned}\Pi^{00}(t) = {} & \Psi(T,t)F\tilde{\Phi}^{00}(T,t) + \int_t^T \Psi(r,t)Q(r)\tilde{\Phi}^{00}(r,t)\,dr \\ & + \int_t^T \Psi(r,t)\Pi^{00}(r)R(r)[\Pi^{00}(r)\tilde{\Phi}^{00}(r,t) + \Pi^{01}(r)\tilde{\Phi}^{10}(r,t)]\,dr,\end{aligned} \tag{5.38}$$

$$\begin{aligned}\Pi^{01}(t) = {} & \Psi(T,t)F\tilde{\Phi}^{01}(T,t) + \int_t^T \Psi(r,t)Q(r)\tilde{\Phi}^{01}(r,t)\,dr \\ & + \int_t^T \Psi(r,t)\Pi^{00}(r)R(r)[\Pi^{00}(r)\tilde{\Phi}^{01}(r,t) + \Pi^{01}(r)\tilde{\Phi}^{11}(r,t)]\,dr,\end{aligned} \tag{5.39}$$

$$\Pi^{10}(t) = \Pi^{01}(t)^*, \tag{5.40}$$

$$\begin{aligned}\Pi^{11}(t) = {} & \tilde{\Phi}^{01}(T,t)^*F\tilde{\Phi}^{01}(T,t) + \int_t^T \tilde{\Phi}^{01}(r,t)^*Q(r)\tilde{\Phi}^{01}(r,t)\,dr \\ & + \int_t^T [\Pi^{00}(r)\tilde{\Phi}^{01}(r,t) + \Pi^{01}(r)\tilde{\Phi}^{11}(r,t)]^*R(r)[\Pi^{00}(r)\tilde{\Phi}^{01}(r,t) \\ & + \Pi^{01}(r)\tilde{\Phi}^{11}(r,t)]\,dr,\end{aligned} \tag{5.41}$$

where $\tilde{\Phi}(t, s)$, *defined by* (5.16), *and* $\Psi(t, s)$, *defined by* (5.21), *only depend on* Π^{00} *and* Π^{01}.

Proof. To derive (5.38) and (5.39) we decompose (5.19) and use the fact that $\Pi^0(t) = P^0(t)$, and to derive (5.41) we decompose (5.17). ▧

Remark. Equation (5.38) relates Π^{00} to Π^{00} and Π^{01}, equation (5.39) relates Π^{01} to Π^{00} and Π^{01} and equation (5.41) explicitly relates Π^{11} to Π^{00} and Π^{01}.

Acknowledgment. It is a pleasure to acknowledge interesting and enlightening discussions we have had with Professor J. L. Lions of the University of Paris.

Notes added in proof.

1. It can be shown that in Proposition 5.4 and Corollary 5.5 the map $s \mapsto \Pi(s) : [0, T] \to \mathscr{L}(M^2)$ is continuous and for all $h, \bar{h}$ in $\mathscr{D}$ the map $s \mapsto (h|\Pi(s)\bar{h})$ is in $AC^1(0, T; R)$. Somewhat similar remarks can be made for the map $t \mapsto P^0(t)$ in Proposition 5.6 and Corollary 5.7.

2. The relationship between controllability, stabilizability and the infinite time quadratic cost problem has been clarified. See:

(a) M. C. DELFOUR AND S. K. MITTER, *L^2-stability, stabilizability and the infinite time quadratic cost problem for linear autonomous hereditary differential systems*, Rep. C.R.M.-132, Centre de Recherches Mathématiques, Université de Montréal, 1971; submitted to this Journal.

(b) H. F. VANDEVENNE, *Qualitative properties of a class of infinite dimensional systems*, Doctoral thesis, Electrical Engineering Dept., M.I.T., Cambridge, Mass., 1972.

3. The following reference which appears to be relevant to the present work was pointed out to us by the referee:

A. MANITIUS, *Optimum control of linear time-lag processes with quadratic performance indices*, Proc. 4th IFAC Congress, Warsaw, Poland, 1969.

REFERENCES

[1] Y. ALEKAL, P. BRUNOVSKY, D. H. CHYUNG AND E. B. LEE, *The quadratic problem for systems with time delay*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 673–687.

[2] A. F. BUCKALO, *Explicit conditions for controllability of linear systems with time lag*, Ibid., AC-13 (1968), pp. 193–195.

[3] D. H. CHYUNG AND E. B. LEE, *Linear optimal control problems with time delays*, this Journal, 4 (1966), pp. 548–575.

[4] G. DA PRATO, *Equations d'évolution dans des algèbres d'opérateurs et application à des équations quasi-linéaires*, J. Math. Pures Appl., 48 (1969), pp. 59–107.

[5] M. C. DELFOUR, *Function spaces with a projective limit structure*, J. Math. Anal. Appl., to appear.

[6] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delays. I—General case*, Centre de Recherches Mathématiques, Université de Montréal, Canada, 1970; J. Differential Equations, to appear.

[7] ———, *Hereditary differential systems with constant delays. II—A class of affine systems and the adjoint problem*, Centre de Recherches Mathématiques, Université de Montréal, Canada, 1970; submitted to J. Differential Equations.

[8] ———, *Systèmes d'équations différentielles héréditaires à retards fixes. Théorèmes d'existence et d'unicité*, C. R. Acad. Sci. Paris, 272 (1971), pp. 382–385.

[9] ———, *Systèmes d'équations différentielles héréditaires à retards fixes. Une classe de systèmes affines et le problème du système adjoint*, Ibid., 272 (1971), pp. 1109–1112.

[10] ———, *Le contrôle optimal des systèmes gouvernés par des équations différentielles héréditaires affines*, Ibid., 272 (1971), pp. 1715–1718.

[11] ———, *Controllability and observability for infinite-dimensional systems*, this Journal, 10 (1972).

[12] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. I*, Interscience, New York, 1967.

[13] D. H. ELLER, J. K. AGGARWAL AND H. T. BANKS, *Optimal control of linear time-delay systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 678–687.

[14] H. O. FATTORINI, *Some remarks on complete controllability*, this Journal, 4 (1966), pp. 686–694.

[15] ———, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.

[16] J. K. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1971.

[17] A. HALANAY, *On a boundary-value problem for linear systems with time lag*, J. Differential Equations, 2 (1966), pp. 47–56.

[18] ———, *On the controllability of linear difference-differential systems*, Mathematical Systems Theory and Economics, II, H. W. Kuhn and G. P. Szegö, eds., Springer-Verlag, Berlin, 1970, pp. 329–335.

[19] D. R. HALEY, *R^n-Controllability for differential-delay equations*, Doctoral thesis, Mathematics Dept., University of Maryland, College Park, 1970.

[20] V. JURDJEVIC, *Abstract control systems: Controllability and observability*, this Journal, 8 (1970), pp. 424–439.

[21] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[22] F. M. KIRILLOVA AND S. V. CHURAKOVA, *The problem of the controllability of linear systems with an after-effect*, Differential Equations, 3 (1967), pp. 221–225.

[23] N. N. KRASOVSKII, *On the analytic construction of an optimal control in a system with time lags*, Prikl. Mat. Meh., 26 (1962), pp. 39–51 = J. Appl. Math. Mech., 26 (1962), pp. 50–67.

[24] ———, *Optimal processes in systems with time lags*, Proc. 2nd IFAC Conf., Basel, Switzerland, 1963.

[25] H. J. KUSHNER AND D. I. BARNEA, *On the control of a linear functional-differential equation with quadratic cost*, this Journal, 8 (1970), pp. 257–272.

[26] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968; English transl., Springer-Verlag, Berlin, New York, 1971.

[27] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[28] S. K. MITTER, *Optimal control of distributed parameter systems*, Joint Automatic Control Conference (Boulder, Colo., 1969), ASME, New York, 1969, pp. 13–48.

[29] V. M. POPOV, Private communication.

[30] D. W. ROSS AND I. FLÜGGE-LOTZ, *An optimal control problem for systems with differential-difference equation dynamics*, this Journal, 7 (1969), pp. 609–623.

[31] L. M. SILVERMAN AND H. E. MEADOWS, *Controllability and observability in time-variable linear systems*, this Journal, 5 (1967), pp. 64–73.

[32] G. S. TAHIM, *Controllability of linear systems with time lag*, Proc. 3rd Allerton Conf. on Circuit and System Theory, Urbana, Ill., 1965, pp. 346–353.

[33] L. WEISS, *On the controllability of delay-differential systems*, this Journal, 5 (1967), pp. 575–587.

[34] ———, *An algebraic criterion for controllability of linear systems with time delay*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 443–444.

[35] ———, *New aspects and results in the theory of controllability for various system models*, Joint Automatic Control Conference (Boulder, Colo., 1969), ASME, New York, 1969, pp. 453–456.

[36] ———, *Lectures on controllability and observability*, C.I.M.E. Seminar on Controllability and Observability, Edizioni Cremonese, Rome, 1969, pp. 205–289.

[37] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

CONTROLLABILITY AND OBSERVABILITY FOR INFINITE-DIMENSIONAL SYSTEMS*

M. C. DELFOUR† AND S. K. MITTER‡

Abstract. This paper systematically studies the notions of controllability and observability for an affine abstract system defined in a Hilbert space with initial data, controls and observations also belonging to Hilbert spaces. Necessary and sufficient conditions are obtained in that framework and the duality property is studied. This theory can find applications in the study of "boundary controllability" and "boundary observability" for parabolic partial differential equations. Specific results have been obtained for affine hereditary differential systems defined in the M^2-space framework (cf. Delfour and Mitter [1], [2], [5]).

1. Introduction. This paper systematically studies the notions of controllability and observability for an affine abstract system defined in a Hilbert space with initial data, controls and observations also belonging to Hilbert spaces. Necessary and sufficient conditions are obtained in that framework and the duality property is studied. In fact, the notion of controllability was chosen in such a way that the duality between the two notions reduces to the notions of topological duality and adjoint map. The choice of this framework was motivated by the work of H. O. Fattorini [8], [9], S. K. Mitter [13], V. Jurdjevic [10] and J. L. Lions [12], and the main ideas all arise from the study of partial differential equations.

This theory finds applications in the study of the notions of "boundary controllability" and "boundary observability" for parabolic partial differential equations. But it is in the theory of affine hereditary differential systems (HDS) that the most interesting applications are found. It is well known that the state space of a HDS is infinite-dimensional. When such a system is studied in the framework of the space $M^2(-b, 0; H)$, its state space is a Hilbert space (cf. Delfour and Mitter [1], [2], [3], [4]). Therefore it was possible to adapt techniques developed by J. L. Lions [12] for optimal control problems with a quadratic cost (cf. Delfour and Mitter [5], [6]). It was also possible to use the results of this paper to study the various notions of controllability and observability for affine HDS in the framework of the space $M^2(-b, 0; H)$ (cf. Delfour and Mitter [5]).

Notations and terminology. Given two real linear spaces X and Y and a linear map $T: X \to Y$, the image of T in Y will be denoted by $\mathrm{Im}(T)$ and the kernel of T in X by $\mathrm{Ker}(T)$. Let H and K be two Hilbert spaces and $T: H \to K$ be a continuous linear map. The adjoint of T will be denoted $T^*(\in \mathscr{L}(K, H))$. When $H = K$ we shall write $T \geqq 0$ for a positive operator ($(Tx|x) \geqq 0$ for all x) and $T > 0$ for a positive definite operator ($(Tx|x) > 0, x \neq 0$). The identity map in $\mathscr{L}(H)$ is

* Received by the editors September 13, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Centre de Recherches Mathématiques, Université de Montréal, Montréal 101, Québec, Canada.

‡ Electronic Systems Laboratory and Electrical Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The work of this author was supported in part by the National Science Foundation under Grant GK-25781 and by the Air Force Office of Scientific Research under Grant AFSOR-70-1941.

written I. The restriction of the map $x:[0, \infty[\to X$ to the interval $[0, t]$ is denoted $\pi_t x$ for all $t \in]0, \infty[$.

2. Definitions. Let D, U, X, Y be real Hilbert spaces with norm $|\cdot|$ and inner product $(\cdot|\cdot)$ indexed by the appropriate space. The space D is the *space of data* (or *initial states*), U is the *space of controls*, $L^2_{\text{loc}}(0, \infty; U)$ is the *space of control maps*, X is the *space of evolution* and Y is the *space of observations*. Consider the *affine system* $\mathscr{A}$ at time $t \in [0, \infty[$ which is characterized by the *evolution map*

$$(d, u) \mapsto \phi(t; d, u) = F(t)d + S(t)\pi_t(u) + g(t) \tag{1}$$

defined on $D \times L^2_{\text{loc}}(0, \infty; U)$ with values in X, where

(i) for all $d \in D$, $t \mapsto F(t)d:[0, \infty[\to X$ is continuous;

(ii) $t \mapsto g(t):[0, \infty[\to X$ is continuous;

(iii) $S(0) = 0$, $S(t) \in \mathscr{L}(L^2(0, t; U), X)$ for all $t \in]0, \infty[$;

(iv) for all $u \in L^2_{\text{loc}}(0, \infty; U)$, $t \mapsto S(t)\pi_t(u):[0, \infty[\to X$ is continuous.

Remark. D can be thought of as the *state space* of system $\mathscr{A}$. When $X = D$ the above formulation is similar to Kalman's definition of a dynamical system, where (1) is the *state transition map*. In general (that is, $X \neq D$), (1) is not a state transition map; it only describes the evolution of the system. This occurs in the theory of hereditary differential equations where the space in which the system evolves is not the state space. Finally, the space X should not be confused with the space of observations Y which can be different from both X and D.

For system $\mathscr{A}$ we define an *observer* $\overline{Z}(t)$ *at time* $t \in]0, \infty[$ as an element of $\mathscr{L}(L^2(0, t; X), L^2(0, t; Y))$; the *observation* $\zeta(t, d, u)$ at time t is given by $\zeta(t, d, u) = \overline{Z}(t)\pi_t(\phi(\cdot; d, u))$.

DEFINITION 1. Let $T, 0 < T < \infty$, be fixed.

(i) The data $d \in D$ is *controllable at time* T to a point $x \in X$ if there exists a sequence of control maps $\{u_n\}$ in $L^2(0, T; U)$ such that $\phi(T; d, u_n) \to x$; d is said to be *strictly controllable at time* T to x if there exists a control map u in $L^2(0, T; U)$ such that $\phi(T; d, u) = x$. System $\mathscr{A}$ is said to be *controllable* (*strictly controllable*) *at time* T if all points of D are controllable (strictly controllable) at time T to all points of X.

(ii) Given $u \in L^2(0, T; U)$, a state $d \in D$ is said to be *observable at time* T if d can be uniquely determined from a knowledge of u and the *observation map* $\zeta(T, d, u)$; system $\mathscr{A}$ is said to be *observable at time* T if all states in D are observable at time T.

DEFINITION 2. The data $d \in D$ is *controllable* (*strictly controllable*) *to the origin* if there exists a finite time $T > 0$ for which d is controllable (strictly controllable) at time T to the origin. System $\mathscr{A}$ is said to be *controllable* (*strictly controllable*) *to the origin* if all points of D are controllable (strictly controllable) to the origin.

PROPOSITION 3. *When X is finite-dimensional the notions of controllability and strict controllability in Definitions* 1 *and* 2 *coincide.*

Proof. All subspaces of a finite-dimensional Hilbert space are closed.

In general, the notions of controllability at time T and controllability to the origin are not equivalent. Very often the sufficient conditions for controllability to the origin are obtained in the following way.

PROPOSITION 4. *$\mathscr{A}$ is controllable to the origin if there exists a finite time $T > 0$ for which system $\mathscr{A}$ is controllable at time T.*

In this paper we shall limit our investigation to the problem of controllability at time T. Definition 2 was introduced for completeness.

3. Main results. We first derive necessary and sufficient conditions for the various notions introduced in Definition 1. Prior to our main theorem we need three lemmas and a definition.

LEMMA 5. *Let H and K be Hilbert spaces and let $\{T(t)|t \in [0, \infty[\}$ be a family of elements of $\mathscr{L}(H, K)$. Assume that for all $h \in H$ the map $t \mapsto T(t)h$ is continuous. Then $T \in L^{\infty}_{\text{loc}}(0, \infty; \mathscr{L}(H, K))$.*

Proof. The proof is a straightforward adaptation of the proof of a similar lemma in [7, Lemma 3, p. 616].

DEFINITION 6. Let $T > 0$ be finite. The maps $\bar{F}(T): D \to L^2(0, T; X)$ and $\bar{S}(T): L^2(0, T; U) \to L^2(0, T; X)$ are defined by $(\bar{F}(T)d)(t) = F(t)d$ and $(\bar{S}(T)u)(t) = S(t)\pi_t(u)$.

LEMMA 7. *The maps $\bar{F}(T)$ and $\bar{S}(T)$ are linear and continuous. In particular, the map $(d, u) \mapsto \pi_T(\phi(\cdot\,; d, u))$ (resp. $(d, u) \mapsto \zeta(T, d, u)$) defined in $D \times L^2(0, T; U)$ with values in $L^2(0, T; X)$ (resp. $L^2(0, T; Y)$) is affine and continuous.*

Proof. By Lemma 5, $\bar{F}(T) \in L^{\infty}(0, T; \mathscr{L}(D, X))$. Hence $\|\bar{F}(T)d\|_{L^2} \leqq \|\bar{F}(T)\|_{L^{\infty}} |d|_D$. The proof is identical for $\bar{S}(T)$. The remainder of the lemma is now obvious.

LEMMA 8. *Let H and K be Hilbert spaces and let $\Lambda \in \mathscr{L}(H, K)$ be given. The following statements are equivalent:*

(i) *Λ is injective (resp. has a dense image in K);*

(ii) *Λ^* has a dense image in H (resp. is injective);*

(iii) *$\Lambda^*\Lambda > 0$ (resp. $\Lambda\Lambda^* > 0$).*

Proof. (i) $\Leftrightarrow$ (ii). $\operatorname{Ker} \Lambda = (\operatorname{Im} \Lambda^*)^{\perp} = (\overline{\operatorname{Im} \Lambda^*})^{\perp}$, where $^{\perp}$ denotes the orthogonal complement in H. Hence $\overline{\operatorname{Im} \Lambda^*} = H$ if and only if $\operatorname{Ker} \Lambda = 0$.

(i) $\Leftrightarrow$ (iii). By definition, $\operatorname{Ker} \Lambda^*\Lambda \supset \operatorname{Ker} \Lambda$. Conversely, $h \in \operatorname{Ker} \Lambda^*\Lambda$ implies that $|\Lambda h|_K^2 = (h|\Lambda^*\Lambda h) = 0$. Hence $\operatorname{Ker} \Lambda^*\Lambda \subset \operatorname{Ker} \Lambda$. When Λ is injective, $\Lambda^*\Lambda$ is injective. By symmetry, $\Lambda^*\Lambda > 0$.

THEOREM 9.

(i) *$\mathscr{A}$ is controllable at time T if and only if the image of $S(T)$ is everywhere dense in X.*

(ii) *$\mathscr{A}$ is strictly controllable at time T if and only if the map $S(T)$ is surjective.*

(iii) *$\mathscr{A}$ is observable at time T if and only if the composite map $\bar{Z}(T)\bar{F}(T)$ is injective.*

Proof. (i) Let the image of $S(T)$ be everywhere dense in X. Then for arbitrary $d \in D$ and $x \in X$ there exists a sequence $\{u_n\}$ in $L^2(0, T; U)$ such that $S(T)u_n \to x - F(T)d - g(T)$, that is, $\phi(T; d, u_n) \to x$. Conversely, for arbitrary $y \in X$, let $h = 0$ and $x = g(T) + y$. Since $\mathscr{A}$ is controllable at time T there exists $\{u_n\}$ in $L^2(0, T; U)$ such that $\phi(T; 0, u_n) \to x = g(T) + y$, that is, $S(T)u_n \to y$. Hence the image of $S(T)$ is everywhere dense in X.

(ii) When $S(T)$ is surjective, $\mathscr{A}$ is clearly strictly controllable at time T. Conversely, for arbitrary $y \in X$, let $h = 0$ and $x = g(T) + y$; there exists $u \in L^2(0, T; U)$ such that $g(T) + y = x = S(T)u + g(T)$. Hence $S(T)$ is surjective.

(iii) Let $\bar{Z}(T)\bar{F}(T)$ be injective. For arbitrary $d \in D$ consider the observation $\zeta(T, d, 0)$ and assume that there exists $d' \neq d$ such that $\zeta(T, d, 0) = \zeta(T, d', 0)$. Then $\bar{Z}(T)(\bar{F}(T)(d - d')) = 0$ in contradiction with our initial hypothesis. Conversely, if the system is observable at time T, then for all $d \neq 0$,

$$\bar{Z}(T)(\bar{F}(T)d + \pi_T(g)) \neq \bar{Z}(T)(\pi_T(g)).$$

But this means that $\bar{Z}(T)(\bar{F}(T)d) \neq 0$ and that the composite map $\bar{Z}(T)\bar{F}(T)$ is injective.

COROLLARY 10. *The following conditions are equivalent*:

(i) $\mathscr{A}$ *is controllable* (*observable*) *at time* T;

(ii) $\overline{\mathrm{Im}\,(S(T))} = X$ ($\bar{Z}(T)\bar{F}(T)$ *is injective*);

(iii) $(S(T))^*$ *is injective* ($\overline{\mathrm{Im}\,(F(T)^*Z(T)^*)} = D$);

(iv) $S(T)(S(T))^* > 0$ ($\bar{F}(T)^*\bar{Z}(T)^*\bar{Z}(T)\bar{F}(T) > 0$).

Remarks. (i) The evolution space X is the direct sum of the closed linear subspace $X_c = \overline{\mathrm{Im}(S(T))}$ and its orthogonal complement X_u. Similarly, the state space D is the direct sum of the closed linear subspace $D_u = \mathrm{Ker}(\bar{Z}(T)\bar{F}(T))$ and its orthogonal complement D_0 which is isomorphic to the quotient D/D_u. There exist linear maps

$$\mathscr{F}(T): D_0 \oplus D_u \to X_c \oplus X_u \quad \text{and} \quad \mathscr{S}(T): L^2(0, T, U) \to X_c \oplus X_u$$

such that system $\mathscr{A}$ at time T is equivalent to the following canonical system:

$$(d, u) \mapsto \mathscr{F}(T)d + \mathscr{S}(T)u : (D_0 \oplus D_u) \times L^2(0, T; U) \to X_1 \oplus X_2.$$

(ii) X_c (resp. X_u) is usually referred to as the *controllable* (resp. *uncontrollable*) part of $\mathscr{A}$. Similarly, D_0 (resp. D_u) is referred to as the *observable* (*unobservable*) part of $\mathscr{A}$.

(iii) When $\mathscr{A}$ is strictly controllable at T, $S(T)$ is surjective. Thus it is a topological isomorphism by the open-mapping theorem. In particular, when a point $h \in D$ is strictly controllable to a point $x \in X$, all points in a neighborhood of h are strictly controllable to points in a neighborhood of x.

The duality between the notions of observability and controllability is a consequence of Lemma 8.

DEFINITION 11. (i) The *controlled dual system* $\mathscr{A}_c^*$ of $\mathscr{A}$ is defined at time $t \in [0, T]$ by the map

$$y \mapsto \chi(t; y) = \int_t^T F(s)(\bar{Z}(T)^*y)(s)\, ds : L^2(0, T; Y) \to D.$$

(ii) The *observed dual system* $\mathscr{A}_0^*$ of $\mathscr{A}$ is defined by the map

$$x \mapsto \psi(x) = S(T)^*x : X \to L^2(0, T; U).$$

Remark. For dual systems, Y is the space of controls, D is the space of evolution, X is the space of data and U is the space of observations.

DEFINITION 12. Assume that the following hypotheses are satisfied:

(i) the operator $S(T)$ has the integral representation

$$S(T)u = \int_0^T R(T, t)B(t)u(t)\, dt,$$

where $B \in L^\infty(0, T; \mathscr{L}(U, D))$, $R(T, t) \in \mathscr{L}(D, X)$ and the map $t \mapsto R(T, t)d : [0, T] \to X$ is continuous for all d in D;

(ii) there exists $Z \in L^\infty(0, T; \mathscr{L}(X, Y))$ such that

$$(\bar{Z}(T)x)(t) = Z(t)x(t) \quad \text{a.e. in } [0, T] \quad \text{for all } x \in L^2(0, T; X).$$

The simultaneously controlled and observed adjoint system $\mathscr{A}^*$ is defined as follows:

$$\phi^*(t; x, y) = R(T, t)^* x + \int_t^T F(s)^* Z(s)^* y(s)\, ds \quad \text{(evolution map)},$$

$$\zeta^*(t; x, y) = B(t)^* \phi^*(t; x, y) \quad \text{(observation map)}.$$

Remarks. (i) Notice that by Lemma 8, the maps $t \mapsto R(T, t) : [0, T] \to \mathscr{L}(D, X)$ and $s \mapsto F(s)^* : [0, T] \to \mathscr{L}(X, D)$ are in $L^\infty(0, T; \mathscr{L}(D, X))$ and $L^\infty(0, T; \mathscr{L}(X, D))$. Thus the above hypotheses make sense. Moreover, when $x = 0$ system $\mathscr{A}^*$ coincides with $\mathscr{A}_c^*$ and when $y = 0$ it coincides with $\mathscr{A}_0^*$.

(ii) For system $\mathscr{A}^*$ the direction of time has been reversed and we must speak of controllability and observability at time 0 (zero) instead of at time T.

THEOREM 13. *System $\mathscr{A}$ is controllable (resp. observable) at time T if and only if system $\mathscr{A}^*$ is observable (resp. controllable) at time* 0.

Proof. The proof follows from Definitions 11 and 12, Lemma 8 and Corollary 10.

REFERENCES

[1] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delays. I—General case*, Centre de recherches mathématiques, Université de Montréal, Canada, 1970; submitted to J. Differential Equations.

[2] ———, *Hereditary differential systems with constant delays. II—A class of affine systems and the adjoint problem*, Centre de recherches mathématiques, Université de Montréal, Canada, 1970; submitted to J. Differential Equations.

[3] ———, *Systèmes d'équations différentielles héréditaires à retards fixes. Théorèmes d'existence et d'unicité*, C.R. Acad. Sci. Paris, 272 (1971), pp. 382–385.

[4] ———, *Systèmes d'équations différentielles héréditaires à retards fixes. Une classe de systèmes affines et le problème du système adjoint*, Ibid., 272 (1971), pp. 1109–1112.

[5] ———, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972).

[6] ———, *Le contrôle optimal de systèmes gouvernés par des équations différentielles héréditaires affines*, C.R. Acad. Sci. Paris, 272 (1971), pp. 1715–1718.

[7] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. I*, Interscience, New York, 1967.

[8] H. O. FATTORINI, *Some remarks on complete controllability*, this Journal, 4 (1966), pp. 686–694.

[9] ———, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.

[10] V. JURDJEVIC, *Abstract control systems: Controllability and observability*, this Journal, 8 (1970), pp. 424–439.

[11] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[12] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968; English transl., Springer-Verlag, Berlin, New York, 1971.

[13] S. K. MITTER, *Optimal control of distributed parameter systems*, Joint Automatic Control Conference (Boulder, Colo., 1969), ASME, New York, 1969, pp. 13–48.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

THE MAXIMUM PRINCIPLE AND CONTROLLABILITY OF NONLINEAR EQUATIONS*

JAMES A. YORKE†

Abstract. The main result proved is that a nonlinear control equation is controllable if a related linear equation is controllable. The result allows the set of control values to be discrete and it is not assumed that "small" values of the control are available. The methods used are closely related to the Pontryagin maximum principle.

Let Ω be a subset of R^m and let $f: R \times R^n \times R^m \to R^n$. Assume $f(t, x, u)$ and $f_x(t, x, u)$ are continuous in (t, x, u). Consider the nonlinear differential equation

$$\text{(NL)} \qquad x'(t) = f(t, x(t), u(t)) \qquad \left(' = \frac{d}{dt}\right),$$

where $u(t)$ is a bounded measurable function with values in Ω.

For t_0 and t_1, $t_0 < t_1$, in R define the reachable set $A_{\text{NL}} = A_{\text{NL}}(t_0, t_1)$ to be the set of all points $x(t_1)$, where $x(t)$ is a solution of (NL) with $x(t_0) = 0$, considering all bounded measurable control functions $u: [t_0, t_1] \to \Omega$. Thus, $A_{\text{NL}}(t_0, t_1)$ is the set of points which can be "reached" at time t_1 from 0 by using appropriate control functions.

We shall compare controllability of (NL) with controllability for the linear equation

$$\text{(L}_Q\text{)} \qquad x'(t) = L(t)x(t) + v(t), \qquad v(t) \in Q(t) \subset R^n,$$

where v is a measurable function, $Q(t)$ is a convex set for each t, and where $L(t)$ is $f_x(t, 0, 0)$, an $n \times n$ matrix which is continuous in t. For a set $\Lambda \subset R^n$, let $K(\Lambda)$ be the unbounded closed convex cone of Λ; that is, $K(\Lambda)$ is the smallest closed convex set containing Λ such that

$$v \in K(\Lambda) \quad \text{implies} \quad \alpha v \in K(\Lambda) \quad \text{for all } \alpha \geqq 0.$$

Let $f(t, 0, \Omega)$ denote $\{f(t, 0, u): u \in \Omega\}$. Let the times t_0 and t_1 be fixed. Our principal result is the following theorem.

THEOREM 1. *Let $\Lambda(t)$ be the set $f(t, 0, \Omega)$ and $Q(t) = K(\Lambda(t))$. Assume 0 is in Ω and $f(t, 0, 0) = 0$. If 0 belongs to the interior of the reachable set A_{L_Q} for the linear equation* (L$_Q$), *then 0 belongs to the interior of the reachable set A_{NL} for the nonlinear equation* (NL).

Theorem 1 says that the reachable set for a nonlinear equation contains a neighborhood of 0 whenever the reachable set (using the same t_0 and t_1) of a certain related linear equation contains a neighborhood of 0. Criteria for controllability for linear systems can be found in [2]. For autonomous linear controllability results for which 0 is on the boundary of the convex hull of Ω, see [5]. For non-

* Received by the editors June 8, 1970, and in revised form April 1, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, College Park, Maryland 20740. This work was supported in part by the National Science Foundation under Grant GP-17601.

autonomous linear controllability results, see [1]. Of particular interest is the case in which Ω is a finite set and f is independent of t. Markus (see [3] or [2, p. 373]) proved a controllability result for time-independent nonlinear equations by writing $f(x, u)$ as

$$f(x,u) = \frac{\partial}{\partial x} f(0,0)x + \frac{\partial}{\partial u} f(0,0)u + o(|x| + |u|), \tag{1}$$

where $f(0, 0) = 0$. His result is quite different in spirit from ours. Control values which are not close to $u = 0$ are of no value to him since he represents f by (1); that is, his theorem involves a representation of f which is useful only for considering small values of x and u. It is certainly reasonable to consider x small since our theorems discuss some neighborhood of $x = 0$. In many situations, particularly if bang-bang control is desirable, it is helpful to be able to use control values u for which $f(x, u)$ is either 0 or large and not have to rely on having u (and hence $f(x, u)$) almost 0.

Examples. We consider two cases of controlling the pendulum,

$$y''(t) = -\sin y(t) + g(u(t)), \qquad u(t) \in \Omega \subset R,$$

where $g: R \to R$ with $g(0) = 0$ and $0 \in \Omega$.

Let $x = (x_1, x_2)$, where $x_1 = y$, $x_2 = y'$, so that the pendulum equation becomes

$$x' = \begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} x_2 \\ -\sin x_1 + g(u) \end{pmatrix} = f(x,u). \tag{P}$$

Clearly,

$$L = L(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad Q(t) = \{0\} \times K(g(\Omega)),$$

since $f(0, \Omega) = \{0\} \times g(\Omega)$. We may write (L_Q) as

$$x'(t) = Lx(t) + v(t), \qquad v(t) \in Q(t).$$

Notice that this system is controllable when $Q(t) = \{0\} \times R$.

Case 1. Let $\Omega = \{-b, 0, b\}$ for some $b > 0$, and assume $g(-b) < 0 < g(b)$. Then $K(g(\Omega)) = R$. From the controllability, it follows that $A_{L_Q} = R^2$ and from Theorem 1, (0, 0) belongs to the interior of A_P (for any $t_1 > t_0$).

Case 2. Let $\Omega = \{0, b\}$, where $b \neq 0$ and $g(b) \neq 0$. Then $K(g(\Omega))$ is either $[0, \infty)$ or $(-\infty, 0]$. In this case, it can be shown easily [5] that 0 is in the interior of A_{L_Q} if and only if $t_1 > t_0 + \pi$. It follows from Theorem 1 that 0 is in the interior of A_P if $t_1 - t_0 > \pi$. (It can be shown that $t_1 - t_0 > \pi$ is a necessary and sufficient condition.)

Lemma 1 shows that $\Lambda(t)$ can be replaced in linear equations by the unbounded closed convex cone $K(\Lambda(t))$. A consideration of the cases' dimension $m = 1$ or 2 shows that $K(\Lambda(t))$ is frequently geometrically much simpler than $\Lambda(t)$.

LEMMA 1. *Let $L(t)$ be an $n \times n$ matrix which is continuous in t. Let $\Lambda(t) \subset R^n$ be the continuous image of a set $\Omega \subset R^m$ with $0 \in \Lambda(t)$ for all t (as, for example, $\Lambda(t) = f(t, 0, \Omega)$). Let $Q(t) = K(\Lambda(t))$, and consider*

$$x'(t) = L(t)x(t) + v(t), \qquad v(t) \in \Lambda(t) \subset R^n. \tag{L_Λ}$$

Then 0 is interior to $A_{L_\Lambda}(t_0, t_1)$ if and only if 0 is interior to $A_{L_Q}(t_0, t_1)$.

Remark. Actually, since $Q(t)$ is an unbounded cone, so is $A_{L_Q}(t_0, t_1)$, and 0 is interior to A_{L_Q} if and only if A_{L_Q} equals R^n.

We postpone the proof of this lemma.

The proof in [2] of the maximum principle is based on the following lemma [2, p. 252, Lemma 2], which is restated here in our terminology. The solution $\bar{x}(t)$ in [2] is taken to be the zero solution. Let $L(t) = f_x(t, 0, 0)$ and $\Lambda(t) = f(t, 0, \Omega)$.

LEMMA 2. *Let $b(\neq 0)$ be a vector interior to A_{L_Λ}. Then, there exists an open convex set C with $0 \in \bar{C}$ (C closure) such that:*

(i) *for some $\delta > 0$, δb is interior to C, and*

(ii) $C \subset A_{NL}$.

This lemma is given in [2] for time-independent systems. (For the purposes of proving the maximum principle, that case suffices since time-dependent systems are later introduced by means of the transversality conditions.) However, with numerous (but minor) changes, this lemma can be proved for the time-dependent case.

Define the *σ-hull for b*:

$$C_\sigma(b) = \left\{ x \in R^n : 0 < |x| < \sigma \text{ and } \left| \frac{b}{|b|} - \frac{x}{|x|} \right| < \sigma \right\}.$$

If $b \neq 0$ is interior to A_{L_Λ} and C is an open convex set with $0 \in \bar{C}$ satisfying (i) and (ii), then, for all σ sufficiently small, $C_\sigma(b) \subset C$ since A_{L_Λ} is convex. Hence, the conclusion that "there exists an open convex set C with $0 \in \bar{C}$ such that (i) and (ii)" may be replaced by:

(iii) for some $\sigma > 0$, $C_\sigma(b) \subset A_{NL}$.

Of course for any $\sigma > 0$ there is always a $\delta > 0$ with δb interior to $C_\sigma(b)$.

Proof of Theorem 1. Assume the hypotheses of the theorem are true but the conclusion is false. In particular, assume

(iv) 0 is in the interior of $A_{L_Q}(t_0, t_1)$,

(v) 0 is *not* in the interior of $A_{NL}(t_0, t_1)$.

We shall thereby get a contradiction. Notice that 0 is in $A_{NL}(t_0, t_1)$, since $x(t) \equiv 0$ is a solution of (NL). Therefore, (8) implies there exists a sequence $\{x_m\}_1^\infty \subset R^n$ such that $x_m \to 0$ as $m \to \infty$ and no x_m is in A_{NL} (so $x_m \neq 0$). Replacing $\{x_m\}$ by a subsequence if necessary, we may assume $x_m/|x_m|$ is convergent. Let $b_0 = \lim_{m \to \infty} x_m/|x_m|$ (so $|b_0| = 1$). From Lemma 1, (iv) implies

(vi) 0 is in the interior of A_{L_Λ}.

Hence, for some $\lambda > 0$, $b = \lambda b_0$ is interior to A_{L_Λ}. From Lemma 2, (iii) is satisfied. There is an integer M such that for $m > M$, $|x_m| < \sigma$ and $|(b/|b|) - x_m/|x_m|| < \sigma$, so $x_m \in C_\sigma(b)$ for $m > M$. But, "$C_\sigma(b) \subset A_{NL}$" contradicts the assumption that $x_m \notin A_{NL}$. Therefore, no such sequence $\{x_m\}$ can exist and 0 is interior to A_{NL}.

Proof of Lemma 1. $A_{L_\Lambda}(t_0, t_1)$ and $A_{L_Q}(t_0, t_1)$ are convex sets, and $A_{L_\Lambda} \subset A_{L_Q}$ since $Q = K(\Lambda)$, and therefore $\Lambda \subset Q$. (The convexity of A_{L_Λ} follows from the Lyapunov convexity theorem.) Since they are convex, the problem of whether or not they are neighborhoods of 0 is equivalent to the opposite problem of whether or not there is a hyperplane through 0 such that each A lies on one of its two sides. Therefore "0 is interior to A_{L_Λ}" is equivalent to "there exists no η_1

(except 0) such that

$$\eta_1 \cdot y \leqq 0 \quad \text{for all} \quad y \in A_{L_\Lambda}."$$

This is equivalent to (P_Λ): "there exists no η_1 (except 0) such that the function $\eta(t)$ given by

$$\eta'(t) = -L(t)^T \eta(t), \qquad \eta(t_1) = \eta_1 \tag{2}$$

satisfies for almost all $t \in [t_0, t_1]$ the property that

$$\eta(t) \cdot v \leqq 0 \quad \text{for all} \quad v \in \Lambda(t)." \tag{3}$$

Note. This equivalence follows in the usual way by letting $x(t)$ be any solution of (L_Λ) with $x(t_0) = 0$ and computing

$$\begin{aligned}
\eta(t_1) \cdot x(t_1) &= \eta(t_1) \cdot x(t_1) - \eta(t_0) \cdot x(t_0) \\
&= \int_{t_0}^{t_1} \frac{d}{dt}[\eta(t) \cdot x(t)]\, dt \\
&= \int_{t_0}^{t_1} (\eta'(t) \cdot x(t) + \eta(t) \cdot x'(t))\, dt \\
&= \int_{t_0}^{t_1} (-[L(t)^T \eta(t)] \cdot x(t) + \eta(t) \cdot L(t)x(t) + \eta(t) \cdot v(t))\, dt \\
&= \int_{t_0}^{t_1} \eta(t) \cdot v(t)\, dt.
\end{aligned}$$

Since $0 \in \Lambda(t)$ for all t, it follows that "$\eta(t_1) \cdot x(t_1) \leqq 0$ for all solutions having $x(t_0) = 0$" is equivalent to "$\eta(t) \cdot v(t) \leqq 0$ for almost all t for all choices of $v(t)$"; that is, it is equivalent to (3), where η satisfies (2). By the definition of $Q(t)$, (3) is equivalent to

$$\eta(t) \cdot v \leqq 0 \quad \text{for all} \quad v \in Q(t). \tag{4}$$

Therefore, (P_Λ) is equivalent to (P_Q): "there exists no η_1 (except 0) such that the function $\eta(t)$ given by (2) satisfies (4)." This is equivalent to: "0 is interior to A_{L_Q}."

Remark. Theorem 1 is stated assuming u is allowed to be any measurable function $u:[t_0, t_1] \to \Omega$; however, the result remains true if we restrict u always to be piecewise continuous, or in fact if we restrict u to be in any "admissible class" of functions (in the sense of [4]). The restriction to admissible classes is needed in the proof of Lemma 2.

Problems not answered.

(i) We have considered only the class of integrable control functions $u(t)$, although we have pointed out that any "admissible" class (in the sense of [4]) would be acceptable, since we can use any class of functions for which the maximum principle is true. Suppose Ω is convex. Is our Theorem 1 and is the maximum principle true for the class C of continuous (or C^1 or C^∞, etc.) functions $u(t)$? The set C is not an "admissible" class, but it seems likely that the maximum principle holds for this class anyway. In essence, does Lemma 2 remain true?

(ii) Suppose (instead of assuming 0 is in Λ) we were to assume that 0 is in the convex hull of Λ. Suppose, for example, f is independent of t. Then, for some m between 1 and n, there are $\alpha_1, \cdots, \alpha_m$ (which $\alpha_i \geqq 0$ and $\sum_1^m \alpha_i = 1$) and $u_1, \cdots, u_m$ such that 0 is $\sum_1^m \alpha_i f(0, u_i)$. Would Theorem 1 be true if we let L be $\sum_1^m \alpha_i (\partial/\partial x) f(0, u_i)$? Then $X(t) \equiv 0$ could be considered a chattering (or sliding regime) solution of (NL). What kind of maximum principle could be stated for such problems?

(iii) A more powerful tool than Theorem 1 would be provided if we could replace $L(t)$ in (L_Q) by $L(t, u) = (\partial/\partial x) f(t, 0, u)$ which depends on u. There are problems in which (L_Q) is not controllable but (NL) is controllable, for example, if $(\partial/\partial x) f(t, 0, 0) \equiv 0$ is not interior to the convex hull of $f(t, 0, \Omega)$. In such cases, we have to examine $(\partial/\partial x) f(t, 0, u)$ in some way.

REFERENCES

[1] L. Weiss, *Lectures on controllability and observability*, Seminar on Controllability and Observability, Centro Internazionale Matematico Estivo, Rome, 1969, pp. 205–287.

[2] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[3] L. Markus, *Controllability of nonlinear processes*, this Journal, 3 (1965), pp. 78–90.

[4] L. S. Pontryagin, V. G. Boltyanski, R. V. Gamkrelidze and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[5] S. Saperstone and J. A. Yorke, *Controllability of linear oscillatory systems using positive controls*, this Journal, 9 (1971), pp. 253–262.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

CONTROLLABILITY IN LINEAR AUTONOMOUS SYSTEMS WITH POSITIVE CONTROLLERS*

ROBERT F. BRAMMER†

Abstract. This paper presents results on the controllability of the autonomous linear control system in R^n, $\dot{x} = Ax + Bu$, where $u \in \Omega \subset R^m$ without the assumption that the origin in R^m is interior to Ω. Necessary and sufficient conditions are given for null-controllability (controllability of each point in some neighborhood of the origin to the origin) and global null-controllability with uniformly bounded controllers. This paper extends some results of Saperstone and Yorke who considered the problem of the controllability of the above system with $m = 1$ and $\Omega = [0, 1]$ and obtained necessary and sufficient conditions for controllability for this system. Corollaries to the main result include existence of time-optimal controllers and controllability of nonlinear systems. An example of control of an economic system is presented.

1. Introduction. Much work has been devoted to the following control system in R^n,

$$\dot{x} = Ax + Bu, \tag{1.1}$$

where x is an n-vector, A is a real constant $n \times n$ matrix, B is a real constant $n \times m$ matrix, and u is a member of a set of measurable m-vector functions defined on some interval of the real line. The set of these functions, U, is called the set of admissible control functions, and the range space (also called the control restraint set) for the functions is some nonempty subset of R^m. The convex hull of Ω, $\mathrm{CH}(\Omega)$, is defined as the intersection of all convex sets in R^m which contain Ω. A variety of assumptions may be made concerning the nature of the set Ω (compactness, closedness, etc.), but, for the major theorems in linear control theory, $\mathrm{CH}(\Omega)$ is required to have nonempty interior in R^m with the origin contained in this interior.

The system (1.1) with restraint set Ω is called *completely controllable* if, for each pair of points x_0 and x_1 in R^n, there exists a bounded admissible control, $u(t)$, defined on some finite interval $0 \leqq t \leqq t_1$, which steers x_0 to x_1. Specifically, the solution to (1.1), $x(t, u(\cdot))$, satisfies the boundary conditions $x(0, u(\cdot)) = x_0$ and $x(t_1, u(\cdot)) = x_1$. The system is called *null-controllable* if there exists an open set, V, in R^n which contains the origin and for which any $x_0 \in V$ can be controlled to $x_1 = 0$ in finite time. The system is *globally null-controllable* if V can be taken to be R^n. We shall define the *reachable set at time t*, $R(t)$, to be the set of all points to which the origin can be steered at time t by an admissible controller, and the reachable set, R_∞, to be the union, over positive t, of the sets $R(t)$.

For the system (1.1) we have the following well-known theorem [2, p. 84].

THEOREM 1.1. *Consider the system* (1.1) *with control restraint set* $\Omega \subset R^m$ *containing* $u = 0$ *in its interior. The system is null-controllable if and only if the controllability matrix,* $C(A, B)$, *whose columns are the columns of* $[B, AB, \cdots, A^{n-1}B]$, *has rank n.*

* Received by the editors September 7, 1971, and in revised form October 7, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Goddard Space Flight Center, Greenbelt, Maryland 20770. This work was supported in part by the National Science Foundation under Grant GP-27284 and the Goddard Space Flight Center Three-Quarter Credit Graduate Study Program.

Remark 1.2. We note that, since the rank of $C(A, B)$ is equal to the rank of $C(-A, -B)$, the system

$$\dot{x} = -Ax - Bu \tag{1.2}$$

is null controllable if and only if (1.1) is null-controllable. Thus, $u(t)$ steers (1.2) from x_0 to 0 at $t = t_1$ if and only if $u(t_1 - t)$ steers the solution of (1.1) from 0 to x_0 at $t = t_1$. Thus, the reachable set, R_∞, for (1.1) contains a neighborhood of the origin if and only if (1.1) is null-controllable.

The above theorem is typical of many theorems in linear control theory in the sense that the origin is required to be interior to Ω or, at least, $\mathrm{CH}(\Omega)$. However, controllability questions can still be considered for systems in which 0 is not interior to $\mathrm{CH}(\Omega)$. In a recent paper [4] the system (1.1) with $m = 1$ and $\Omega = [0, 1]$ was considered, and the following theorem was proved.

THEOREM 1.3. *The system* (1.1) *with* $m = 1$ *and* $\Omega = [0, 1]$ *is null-controllable if and only if* $C(A, B)$ *has rank* n *and the matrix* A *has no real eigenvalues* (*such matrices are called oscillatory*).

The authors of [4] also consider (1.1) for $m > 1$ and $\Omega = \prod_1^m [0, 1]$. They show that if $C(A, B)$ has rank n and if A has no real eigenvalues, then (1.1) is null-controllable.

The results of this paper extend those presented in [4] and some further results on controllability presented in [3]. The major theorem on controllability presented in this paper is the following result.

THEOREM 1.4. *Consider the control system* (1.1) *with control restraint set satisfying the following conditions.*

$$\textit{The set } \Omega \textit{ contains a vector in the kernel of } B \tag{1.3}$$

(*i.e., there exists* $u \in \Omega$ *satisfying* $Bu = 0$).

$$\textit{The set } \mathrm{CH}(\Omega) \textit{ has nonempty interior in } R^m. \tag{1.4}$$

The following conditions are necessary and sufficient for the null-controllability of (1.1).

$$\textit{The matrix } C(A, B) \textit{ has rank } n. \tag{1.5}$$

$$\textit{There is no real eigenvector } v \textit{ of } A^T \textit{ satisfying } (v, Bu) \leqq 0 \quad \textit{for all } u \in \Omega. \tag{1.6}$$

(*The parentheses denote the scalar product in* R^n.)

Theorem 1.1 and Theorem 1.3 will be proved as corollaries to Theorem 1.4 in § 3. Further results on the complete controllability, null-controllability, and global null-controllability will also be given in § 3.

The following example is adapted from an economic model of Paul A. Samuelson [5, p. 601]. Let $K(t)$ denote the stock of capital in an industry, and $I(t)$ denote the net investment. Thus, $I(t)$ is the rate of change of $K(t)$, or $I = \dot{K}$. Assume that there is an equilibrium level of capital stock denoted by K_0, and that a level of capital above this equilibrium level leads to "a deceleration in the algebraic rate of investment," and, conversely, a level of capital below this level leads to an increase in the rate of investment in the following way. Let $k = K - K_0$, and thus,

$$\dot{k} = \dot{K} = I, \qquad \dot{I} = -\lambda(K - \bar{K}) = -\lambda k.$$

Let $x = \begin{pmatrix} k \\ I \end{pmatrix}$, and we have

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ -\lambda & 0 \end{pmatrix} x$$

which is clearly an oscillatory system. Suppose that it is desired to stop oscillations in the system, but for various political reasons the government desires to apply controls only in the form of subsidies (positive investment) rather than both subsidies and taxation. These subsidies, G, constitute positive investment which is added to the investment in the industry to give $\dot{k} = I + G$. Thus, we have the modified system

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ -\lambda & 0 \end{pmatrix} x + \begin{pmatrix} 1 \\ 0 \end{pmatrix} G.$$

By Theorem 1.4 it follows that the state of this system can be controlled to the origin (equilibrium) by these subsidies.

2. Preliminary lemmas. The proofs of the major theorems to follow all rely on an asymptotic evaluation of the inner product

$$(v, e^{At}Bu), \tag{2.1}$$

where v is an n-vector, and A, B and u are as in Theorem 1.4. The necessary expression will be developed in a series of lemmas. The first lemma is the Jordan canonical form for matrices over the field of real numbers. This theorem is given in [1, p. 97].

LEMMA 2.1. *Let A be a real $n \times n$ matrix. There exists a nonsingular real matrix Q with the following property: the matrix J defined by $J = QAQ^{-1}$ is a block diagonal matrix with blocks of the following forms:*

$$\begin{bmatrix} \lambda_i & 1 & & & & \\ & \lambda_i & 1 & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \lambda_i & 1 \\ & & & & & \lambda_i \end{bmatrix} \tag{2.2}$$

($k \times k$ block corresponding to an elementary divisor $(\lambda - \lambda_i)^k$),

$$\begin{bmatrix} \begin{pmatrix} 0 & \gamma_k \\ 1 & \beta_k \end{pmatrix} & \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} & & & \\ & \begin{pmatrix} 0 & \gamma_k \\ 1 & \beta_k \end{pmatrix} & \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} & & \\ & & \cdot & & \\ & & & \cdot & \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \\ & & & & \begin{pmatrix} 0 & \gamma_k \\ 1 & \beta_k \end{pmatrix} \end{bmatrix} \tag{2.3}$$

($2l \times 2l$ block corresponding to an elementary divisor $(\lambda^2 - \beta_k\lambda - \gamma_k)^l$ with $\beta_k^2 + 4\gamma_k < 0$).

For our purpose we shall group all blocks of the form (2.2) with the same value of λ_i into one block of the form

$$B_i = \lambda_i I_i + F_i, \tag{2.4}$$

where B_i is an $m_i \times m_i$ matrix, m_i is the multiplicity of the real eigenvalue λ_i, I_i is the $m_i \times m_i$ identity matrix, and F_i is an $m_i \times m_i$ matrix whose only possible nonzero elements are 1's on the superdiagonal. The matrix F_i is nilpotent (i.e., $F_i^j = 0$ for some nonnegative integer $j \leqq m_i$).

Furthermore, we shall group together all blocks of the form (2.3) with the same value of β_j into one block of the form

$$C_j = \begin{bmatrix} D_{j_1} & & \\ & \ddots & \\ & & D_{j_l} \end{bmatrix}, \tag{2.5}$$

where D_{j_q} has the form (2.3) with all D_{j_q} having the same value β_j.

Thus, we see that the matrix J has the form

$$J = \begin{bmatrix} B_1 & & & & & \\ & \ddots & & & & \\ & & B_k & & & \\ & & & C_1 & & \\ & & & & \ddots & \\ & & & & & C_p \end{bmatrix}, \tag{2.6}$$

where the blocks B_i contain the k distinct real eigenvalues and the blocks C_j contain the p distinct values β_j. Note that each β_j corresponds to a conjugate pair of complex eigenvalues having real part $\beta_j/2$. Define $B_{k+j} = C_j$, and define the $k + p$ projection matrices P_i by

$$P_i = \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & I_i & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}, \tag{2.7}$$

where I_i is the identity matrix of dimension equal to that of B_i and I_i appears in the block corresponding to B_i. It is clear that these projections satisfy the following conditions:

$$\sum_{1}^{k+p} P_i = I, \tag{2.8}$$

$$P_iP_j = 0 \quad \text{if } i \neq j \quad \text{and} \quad P_i^2 = P_i \quad \text{for all } i = 1, \cdots, k + p, \tag{2.9}$$

$$P_iJ = JP_i \quad \text{for all } i = 1, \cdots, k + p, \tag{2.10}$$

$$JP_i = (\lambda_i I + N_i)P_i, \tag{2.11}$$

where N_i is nilpotent for all $i = 1, \cdots, k$,

$$JP_i = \begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & B_i & & \\ & & & \ddots & \\ & & & & 0 \end{bmatrix}, \tag{2.12}$$

where B_i is oscillatory and all eigenvalues have real part $\beta_i/2$ for all $i = k + 1, \cdots, k + p$.

LEMMA 2.2. *Let D be a real $l \times l$ matrix satisfying $D = \lambda I + N$, where N is nilpotent. Let $v \neq 0$ be an l-vector. The l-vector function $e^{Dt}v$ has the following asymptotic expression:*

$$e^{Dt}v = t^k e^{\lambda t}(z + f(t)), \tag{2.13}$$

where k is a nonnegative integer, $f(t)$ vanishes as t goes to infinity, and $Dz = \lambda z$ with $z \neq 0$.

Proof. If $Nv = 0$, then let $z = v$. Since $D = \lambda I + N$, we have $Dv = \lambda v$ and $e^{Dt}v = e^{\lambda t}v$. Take $k = 0$ and $f(t) \equiv 0$ to obtain the form (2.13).

If $Nv \neq 0$, let k be the largest integer satisfying $N^k v \neq 0$. The integer k is defined since N is nilpotent, and $1 \leqq k \leqq l$. The function $e^{Dt}v$ is given by

$$e^{Dt}v = e^{\lambda t}\left(I + Nt + \cdots + N^k \frac{t^k}{k!}\right)v.$$

Let $z = N^k v/k!$ and define, for $t > 0$,

$$f(t) = \frac{1}{t^k}\left(I + Nt + \cdots + N^{k-1}\frac{t^{k-1}}{(k-1)!}\right)v.$$

Thus,

$$e^{Dt}v = t^k e^{\lambda t}(z + f(t)).$$

Clearly $f(t)$ vanishes as t goes to infinity and $Dz = (\lambda I + N)z = \lambda z + NN^k v/k! = \lambda z + N^{k+1}v/k! = \lambda z$.

LEMMA 2.3. *Let D be an $l \times l$ oscillatory matrix and let v and w be l-vectors. The inner product $(v, e^{Dt}w)$ has the form*

$$(v, e^{Dt}w) = t^j e^{\rho t}(a(t) + g(t)), \tag{2.14}$$

where $a(t)$ is a sum of sinusoidal terms (i.e., $a(t) = \sum_i h_i \sin(\omega_i t + \theta_i)$), and $g(t)$ vanishes as t goes to infinity. If $a(t)$ is identically zero, then $g(t)$ is identically zero.

Proof. See [4, Lemma 3.2].

LEMMA 2.4. *The inner product* (2.1) *can be written in the form*

$$(v, e^{At}Bu) = \sum_{i=1}^{k} t^{j_i} e^{\lambda_i t}((z_i, Bu) + (f_i(t), Bu)) + \sum_{i=k+1}^{k+p} e^{\rho_i t} \sum_{j=1}^{m} t^{r_{ij}}(a_{ij}(t) + g_{ij}(t))u_j \tag{2.15}$$

with the following conditions being satisfied.

(a) *The* λ_i *are the* k *distinct real eigenvalues of the matrix* A, *and* $\lambda_1 > \lambda_2 > \cdots > \lambda_k$.

(b) $A^T z_i = \lambda_i z_i$.

(c) *If* $z_i = 0$, *then* $f_i(t) \equiv 0$.

(d) *The* j_i *and* r_{ij} *are nonnegative integers.*

(e) *The functions* $f_i(t)$ *and* $g_{ij}(t)$ *vanish as* t *goes to infinity.*

(f) *The functions* $a_{ij}(t)$ *are sums of sinusoid terms as in Lemma* 2.3.

(g) *If* $a_{ij}(t) \equiv 0$, *then* $g_{ij}(t) \equiv 0$.

(h) *The* ρ_i *are the* p *distinct real parts of the complex eigenvalues of* A *and* $\rho_1 > \rho_2 > \cdots > \rho_p$.

(i) *The* u_j *are components of the* m*-vector* $u = \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}$.

(j) *If* (1.4) *and* (1.5) *are satisfied and if* v *is not zero, then there exists* $u \in \Omega$ *for which* (2.1) *is not identically zero.*

Proof. First note that it is sufficient to prove the lemma in the coordinate system in which the matrix A^T has the form J (see (2.6)).

Furthermore, all of the conditions (a) through (i) are statements concerning eigenvalues, eigenvectors, scalar functions, or vector functions which tend to zero. All of these notions are independent of the choice of linear coordinates in R^n. Thus, we can take the matrix A^T in the form (2.6) and we have the projections P_i given in (2.7) induced by A^T.

Using the projections P_i, we have

$$\begin{aligned} e^{A^T t}v &= \sum_{i=1}^{k+p} e^{A^T t}P_i v \\ &= \sum_{i=1}^{k} e^{A^T P_i t}P_i v + \sum_{i=k+1}^{k+p} e^{A^T P_i t}P_i v. \end{aligned} \tag{2.16}$$

Applying (2.11) and Lemma 2.2 to the first sum we have

$$\begin{aligned} \sum_{i=1}^{k} e^{A^T P_i t}P_i v &= \sum_{i=1}^{k} \exp[(\lambda_i I + N_i)P_i t]P_i v \\ &= \sum_{i=1}^{k} \exp[(\lambda_i I + N_i)t]P_i v \\ &= \sum_{i=1}^{k} t^{j_i} e^{\lambda_i t}(z_i + f_i(t)). \end{aligned} \tag{2.17}$$

Multiplying these equations on the left by P_i leads to the relation $P_i z_i = z_i$. Furthermore, from the proof of Lemma 2.2, we have $N_i z_i = 0$. Therefore $A^T z_i = A^T P_i z_i = (\lambda_i I + N_i) P_i z_i = (\lambda_i I + N_i) z_i = \lambda_i z_i$. Therefore, (2.16) becomes

$$e^{A^T t} v = \sum_{i=1}^{k} t^{j_i} e^{\lambda_i t}(z_i + f_i(t)) + \sum_{i=k+1}^{k+p} e^{A^T P_i t} P_i v. \tag{2.18}$$

Placing (2.18) into (j) gives

$$(e^{A^T t} v, Bu) = \sum_{i=1}^{k} t^{j_i} e^{\lambda_i t}((z_i, Bu) + (f_i(t), Bu)) + \sum_{i=k+1}^{k+p} (e^{A^T P_i t} P_i v, Bu). \tag{2.19}$$

This proves (a), (b), (c), and the first parts of (d) and (e).

Now consider the second sum in (2.16):

$$\sum_{i=k+1}^{k+p} (e^{A^T P_i t} P_i v, Bu).$$

Since $Bu = \sum_{j=1}^{m} b_j u_j$, where the b_j are the column vectors of B and the u_j are the components of the vector u, we have the following double sum:

$$\sum_{i=k+1}^{k+p} \sum_{j=1}^{m} (e^{A^T P_i t} P_i v, b_j) u_j. \tag{2.20}$$

Consider the inner product $(e^{A^T P_i t} P_i v, b_j)$. The matrix $A^T P_i$ is not, in general, an oscillatory matrix. However, if $A^T P_i$ is considered as a transformation restricted to the image of R^n under P_i, $A^T P_i$ is oscillatory. Moreover, since it is clear that $(e^{A^T P_i t} P_i v, b_j) = (e^{A^T P_i t} P_i v, P_i b_j)$, the inner product can be considered as being restricted to the subspace. Therefore, from Lemma 2.3,

$$(e^{A^T P_i t} P_i v, b_j) = e^{\rho_i t} t^{r_{ij}} (a_{ij}(t) + g_{ij}(t)). \tag{2.21}$$

Note that all the parameters on the right-hand side depend on both i and j except for ρ_i which depends only on i. This is due to the fact that ρ_i is the only possible real part for an eigenvalue of $A^T P_i$. Therefore,

$$\sum_{j=1}^{m} (e^{A^T P_i t} P_i v, b_j) u_j = e^{\rho_i t} \sum_{j=1}^{m} t^{r_{ij}} (a_{ij}(t) + g_{ij}(t)) u_j, \tag{2.22}$$

and thus, from (2.19),

$$(e^{A^T t} v, Bu) = \sum_{i=1}^{k} t^{j_i} e^{\lambda_i t}((z_i, Bu) + (f_i(t), Bu)) + \sum_{i=k+1}^{k+p} e^{\rho_i t} \sum_{j=1}^{m} t^{r_{ij}} (a_{ij}(t) + g_{ij}(t)) u_j.$$

This concludes the proofs of (d) through (i), and it remains to prove (j).

If $(v, e^{At} Bu)$ is identically zero in t for all $u \in \Omega$, then, by setting $t = 0$, $(v, Bu) = 0$ for all $u \in \Omega$. Furthermore, by successive differentiation and evaluation at $t = 0$, $(v, ABu) = 0, \cdots, (v, A^{n-1} Bu) = 0$ for all $u \in \Omega$. It follows by transposition that $(B^T v, u) = 0, \cdots, (B^T (A^T)^{n-1} v, u) = 0$ for all $u \in \Omega$. Since CH(Ω) has nonempty interior, it follows that each vector $B^T v, \cdots, B^T (A^T)^{n-1} v$ is perpendicular to m

independent vectors, and thus each of the former vectors is zero. Thus it follows that $v^T C(A, B) = 0$, and since v is not zero, $C(A, B)$ cannot have rank n. Thus, if $\mathrm{CH}(\Omega)$ has nonempty interior in R^m and if $C(A, B)$ has rank n ((1.4) and (1.5) respectively), then there exists $u \in \Omega$ such that $(v, e^{At}Bu)$ is not identically zero for all t.

Remark 2.5. Note that if z_i is zero, then $f_i(t) \equiv 0$. Thus the ith term is zero, and does not contribute to the first sum in (2.15). If any z_i is zero, then, for the purposes of the results to follow, we can ignore λ_i. It may happen that all z_i are zero, and this case will be considered separately. Therefore, without loss of generality, we can assume all z_i not zero. In the same way if (2.22) is identically zero for all $t > 0$ and for all $u \in \Omega$, then, by selective choice of u, it follows that a_{ij} and g_{ij} are identically zero for all $j = 1, \cdots, m$. Thus in the theorems to follow, λ_1 refers to the largest real eigenvalue for which the associated eigenvector z in (2.15) is not zero. If all $z_i = 0$, then let $\lambda_i = -\infty$. In the same way ρ_1 refers to the largest real part of the complex eigenvalue for which

$$\sum_{j=1}^{m} t^{k_{ij}}(a_{ij}(t) + g_{ij}(t))u_i \neq 0.$$

If all such terms are zero, let $\rho_1 = -\infty$.

We shall also need the following result from linear control theory [2, p. 164].

LEMMA 2.6. *Consider the linear control process* (1.1) *with compact restraint set* Ω. *The reachable set at time* t, $R(t)$, *is compact and convex. If* Ω *is relaxed to its convex hull,* $\mathrm{CH}(\Omega)$, *and if* $R_{\mathrm{CH}}(t)$ *is the reachable set at time* t *for admissible controllers in* $\mathrm{CH}(\Omega)$, *then* $R_{\mathrm{CH}}(t) = R(t)$.

Finally, we shall require the following results. A function $f: R^1 \to C$ is almost periodic if it is continuous and if for every $\varepsilon > 0$ there exists $L(\varepsilon) > 0$ with the property that in any interval of length $L(\varepsilon)$ there is a number s satisfying $|f(t + s) - f(t)| < \varepsilon$ for all $t \in R^1$. The functions $a_{ij}(t)$ in Lemma 2.4 are examples of almost periodic functions, since linear combinations of almost periodic functions are almost periodic. The following results appear in [3] and will be used in § 3.

LEMMA 2.7. *For any almost periodic function* f, *the limit*

$$M(f) = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(t)\, dt$$

exists. If $M(f) = 0$, *and if* $f(t) \leqq |h(t)|$, *where* h *is any function that vanishes as* t *goes to infinity, then* $f(t) \equiv 0$.

LEMMA 2.8. *Let* $f(t)$ *be a real continuous almost periodic function satisfying* $f \neq 0$ *and* $M(f) = 0$. *There exist positive numbers* L *and* q *and a denumerable set of disjoint open intervals* $\{I_j\}$ *for all* $j = 1, \cdots, \infty$ *satisfying*:

(2.23) *If* $t_j \in I_j$ *and* $t_{j+1} \in I_{j+1}$, *then* $t_j < t_{j+1}$.

(2.24) *The length of each interval* I_j *is* L.

(2.25) $$\limsup_{j \to \infty} \{t : t \in I_j\} = \infty.$$

(2.26) $$f(t) > q \quad \text{for all } t \in \bigcup_{j=1}^{\infty} I_j.$$

3. Proof of Theorem 1.4 and related results. We can now use the preliminary results of § 2 to prove Theorem 1.4 and several other related controllability results. First note that $-v$ is an eigenvector of $-A^T$ if and only if v is an eigenvector of A^T, that $(-v, -Bu) \leqq 0$ for all $u \in \Omega$ if and only if $(v, Bu) \leqq 0$ for all $u \in \Omega$, and that $C(-A, -B)$ has rank n if and only if $C(A, B)$ has rank n. Thus, (1.2) satisfies (1.5) and (1.6) if and only if (1.1) satisfies (1.5) and (1.6). Thus, from Remark 1.2, to prove Theorem 1.4 it is necessary and sufficient to show that $R(t)$ contains an open set about the origin for some $t > 0$. Furthermore, by Lemma 2.6, $R(t)$ is a convex set for all $t > 0$, and, by (1.3), if $t_1 < t_2$, then $R(t_1) \subset R(t_2)$ and $0 \in R(t)$ for all $t > 0$. Thus, the reachable set, R_∞, is the union of an increasing collection of convex sets and is, thus, a convex set containing the origin. If the origin is interior to R_∞, then there exist $n + 1$ points in R_∞ whose convex hull contains the origin as an interior point. It follows that these $n + 1$ points must be contained in $R(t)$ for some $t > 0$ since these sets are increasing. Thus, the origin is interior to R_∞ if and only if the origin is interior to $R(t)$ for some $t > 0$. From the above statements we conclude that proving Theorem 1.4 is equivalent to proving that (1.5) and (1.6) are necessary and sufficient to ensure that the origin is interior to R_∞.

If the origin is not interior to R_∞, then there exists a supporting hyperplane through the origin with unit outward normal vector v satisfying $(v, x(t, u(\cdot))) \leqq 0$ for all $t > 0$ and all admissible controllers u. Since $x(t, u(\cdot))$ has the form

$$x(t, u(\cdot)) = \int_0^t e^{A(t-s)} Bu(s)\, ds,$$

the vector v satisfies

$$\left(v, \int_0^t e^{A(t-s)} Bu(s)\, ds\right) \leqq 0.$$

It follows by continuity and a special choice of $u(\cdot)$ that

$$(v, e^{At} Bu) \leqq 0 \quad \text{for all } t > 0 \quad \text{and for all } u \in \Omega. \tag{3.1}$$

Conversely, if the origin is interior to R_∞, it is clear that no nonzero vector v satisfies (3.1). Thus we have proved the following lemma.

LEMMA 3.1. *The system* (1.1) *is controllable if and only if there is no nonzero vector* v *satisfying* (3.1).

Remark 3.2. A similar result appears in [4] but the statement is restricted to oscillatory systems, and the proof is different.

We can now proceed to the proof of Theorem 1.4. The proof will be divided into a proof of necessity and a proof of sufficiency.

Proof of necessity. We first show that (1.5) and (1.6) are necessary for the null-controllability of (1.1). Specifically, we show that if either (1.5) or (1.6) is violated then (3.1) can be satisfied.

Suppose that $C(A, B)$ does not have rank n. There exists a nonzero vector v satisfying $v^T C(A, B) = 0$. It is well known that this implies that $v^T e^{At} B = 0$ (see [2, p. 81]). Therefore, $(v, e^{At} Bu) = 0$ for all $t > 0$ and $u \in \Omega$. Now suppose (1.6) is violated. The inner product $(v, e^{At} Bu) = (e^{A^T} v, Bu) = e^{\lambda t}(v, Bu) \leqq 0$ for

all t and $u \in \Omega$. This proves the necessity of (1.5) and (1.6) for the null-controllability of (1.1).

Proof of sufficiency. We now show that (1.5) and (1.6) are sufficient for the null-controllability of (1.1). Specifically, we show that if (1.5) and (1.6) are satisfied, then no nonzero vector v can satisfy (3.1).

Applying Lemma 2.4 to (3.1) we obtain

$$0 \geqq (v, e^{At}Bu) = \sum_{i=1}^{k} t^{j_i} e^{\lambda_i t}((z_i, Bu) + (f_i(t), Bu)) + \sum_{i=k+1}^{k+p} e^{\rho_i t} \sum_{j=1}^{m} t^{r_{ij}}(a_{ij}(t) + g_{ij}(t))u_j. \tag{3.2}$$

The proof will be divided into three cases:

$$\lambda_1 > \rho_1 \quad \text{or} \quad \lambda_1 = \rho_1 \quad \text{and} \quad j_1 > \max_j r_{1j} \equiv r_1, \tag{3.3}$$

$$\lambda_1 < \rho_1 \quad \text{or} \quad \lambda_1 = \rho_1 \quad \text{and} \quad j_1 < \max_j r_{1j} \equiv r_1, \tag{3.4}$$

$$\lambda_1 = \rho_1 \quad \text{and} \quad j_1 = \max_j r_{1j} \equiv r_1. \tag{3.5}$$

Proof for (3.3). For $t > 0$ divide both sides of (3.2) by $t^{j_1} e^{\lambda_1 t}$ to obtain

$$(z_1, Bu) + F(t) \leqq 0,$$

where $F(t)$ is given by

$$F(t) = (f_1(t), Bu) + \frac{1}{t^{j_1} e^{\lambda_1 t}}\left[\sum_{i=2}^{k} t^{j_i} e^{\lambda_i t}((z_i, Bu) + (f_i(t), Bu)) + \sum_{i=k+1}^{k+p} e^{\rho_i t} \sum_{j=1}^{m} t^{r_{ij}}(a_{ij}(t) + g_{ij}(t))u_j\right].$$

By the assumptions of (3.3), $F(t)$ vanishes as t goes to infinity for all $u \in \Omega$. Therefore $(z_1, Bu) \leqq 0$ for all $u \in \Omega$ which violates (1.6).

Proof for (3.4). For $t > 0$ divide both sides of (3.2) by $t^{r_1} e^{\rho_1 t}$ to obtain

$$\sum_{l_j} (a_{1l_j}(t) + g_{1l_j}(t))u_{l_j} + G(t) \leqq 0.$$

The summation appears because it may happen that $r_1 = r_{1l_1} = r_{1l_2}$ for two distinct indices l_1 and l_2. The function $G(t)$ is analogous to $F(t)$ above and vanishes as t goes to infinity for all $u \in \Omega$. Recall that the functions $a_{ij}(t)$ are sums of sinusoidal terms and hence are almost periodic functions of t.

Thus, if the above inequality is satisfied, we have

$$\sum_{l_j} a_{1l_j}(t)u_{l_j} \leqq |H(t)|,$$

where $H(t) = G(t) + \sum_{lj} g_{1_j}(t)u_j$. Since $H(t)$ vanishes as t goes to infinity for all $u \in \Omega$, it follows from [4, Lemma 4.1] that

$$\sum_{l_j} a_{1l_j}(t)u_{l_j} \equiv 0.$$

It follows, by a special choice of u, that each $a_{1l_j} \equiv 0$, which violates the definition of k_1 (i.e., if all a_{1l_j} are trivial, then the terms involving k_1 are all zero and, thus, would not be considered).

Proof for (3.5). For $t > 0$ divide (3.2) by $t^{j_1} e^{\lambda_1 t} = t^{r_1} e^{\rho_1 t}$ to obtain

$$(z_1, Bu) + \sum_{l_j} a_{1l_j}(t) u_{l_j} + J(t) \leqq 0,$$

where $J(t)$ is analogous to F and H above and vanishes as t goes to infinity for all $u \in \Omega$. The mean value of J exists and is zero as is the mean value of $\sum a_{1l_j}(t) u_{l_j}$ for all $u_{l_j} \in \Omega$. Therefore $(z_1, Bu) \leqq 0$ for all $u \in \Omega$, violating (1.6).

Note that we have assumed that at least one term in the summations of (3.2) is not zero. However, the assumptions (1.3) and (1.4) on Ω and condition (1.5) assure this by the proof of (i) in Lemma 2.4.

COROLLARY 3.3. *Theorem* 1.1 *and Theorem* 1.3 *follow as corollaries.*

Proof. If the origin is interior to Ω, then there is a symmetric neighborhood S of the origin in Ω. Thus, if $(z, Bu) \leqq 0$ for all $u \in \Omega$, then $(z, Bu) = 0$ for all u in S. If z is an eigenvector of A^T, then $e^{\lambda t}(z, Bu) = (e^{A^T t} z, Bu) = (z, e^{A^T} Bu) = 0$ for all $u \in S$. Since S is open in R^m, it follows that $C(A, B)$ does not have rank n. Thus, if $u = 0$ is interior to Ω and if (1.5) is satisfied, then (1.6) is satisfied. Thus, Theorem 1.1 follows from Theorem 1.4.

If $m = 1$ and $\Omega = [0, 1]$, and if A^T has an eigenvector v, then $(v, Bu) = (v, b)u$ where b is the column vector comprising B. Either $(v, b) \leqq 0$ or $(-v, b) \leqq 0$. Therefore, either $(-v, Bu) \leqq 0$ or $(v, Bu) \leqq 0$ for all $u \in \Omega$. Thus, if A has a real eigenvalue, then (1.6) cannot be satisfied. Consequently, Theorem 1.3 follows from Theorem 1.4.

COROLLARY 3.4. *If Ω is a cone with vertex at the origin and has nonempty interior in R^m, then* (1.5) *and* (1.6) *are necessary and sufficient for the reachable set R to be all of R^n.*

Proof. Necessity is clear since without (1.5) or (1.6) no open set about the origin can be reached. To show that (1.5) and (1.6) are sufficient to reach all of R^n, first note that since Ω is a cone with vertex at the origin, if $u(\cdot)$ is an admissible controller, then so is $\alpha u(\cdot)$ for all $\alpha \geqq 0$. Since all the hypotheses of Theorem 1.4 are satisfied, the reachable set is a convex set containing the origin in its interior. Moreover, the reachable set is a cone with vertex at the origin since $x(t, u(\cdot))$ is linear in $u(\cdot)$. Thus the reachable set must be all of R^n.

It is not always necessary to use an unbounded control restraint set in order to reach all of R^n. If the system is not asymptotically stable on any subspace, and thus does not tend to prevent the system from reaching points in R^n with large norm, then a compact restraint set may be sufficient to enable the origin to reach any point in R^n in finite time. The proof of the following theorem extends the results presented in [3, Theorem 5.1 and Corollary 5.2].

THEOREM 3.5. *Consider* (1.1) *with bounded restraint set Ω satisfying* (1.3) *and* (1.4). *Necessary and sufficient conditions that the origin can be steered to any point in R^n in finite time are* (1.5), (1.6) *and that no eigenvalue of A has a negative real part.*

Proof of necessity. We first show that the above conditions are necessary in order to steer the origin to any point in R^n in finite time. Note that (1.5) and (1.6)

are necessary to reach any open set containing the origin, so they are certainly necessary to reach all of R^n.

Clearly, the validity of this theorem is independent of a choice of linear coordinates. Thus, we may assume that the matrix A has the Jordan-type form (2.6) and that one eigenvalue, δ_i, has negative real part. Let y be a vector lying in the eigenspace of δ_i. Thus $P_i y = y$ and $P_j y = 0$ for $j \neq i$. Assume that y can be reached at time T using a controller $u(\cdot)$. Thus,

$$y = \int_0^T e^{A(T-s)}Bu(s)\,ds = \int_0^T e^{As}Bu(T-s)\,ds$$

and

$$y = P_i \int_0^T e^{As}Bu(T-s)\,ds = \int_0^T e^{AP_i s}Bu(T-s)\,ds.$$

Therefore,

$$\|y\| \leqq K\|u\| \int_0^T e^{\delta_i s}|1 + s^j|\,ds,$$

where $\|u\|$ is the essential supremum of $u(\cdot)$. Since $\delta_i < 0$, we have

$$\int_0^T e^{\delta_i s}|1 + s^j|\,ds < \int_0^\infty e^{\delta_i s}|1 + s^j|\,ds < \infty.$$

Thus, the reachable vectors y in this subspace have norms bounded by $C \max \|u\|$, where C is a constant independent of T and u. Thus, if Ω is bounded and if one eigenvalue has a negative real part, then all of R^n cannot be reached.

Proof of sufficiency. To prove that the conditions of the theorem are sufficient to allow any point in R^n to be reached from the origin, assume that y does not lie in the reachable set, R. Since R is convex, there is a hyperplane in R^n through y such that the set R lies on one side of the hyperplane. Let v be the unit outward normal vector to this hyperplane at y. Thus we have the inequality

$$(v, x(t, u(\cdot)) - y) \leqq 0 \tag{3.6}$$

for all positive t and admissible controllers u. Consequently, the inner product

$$(v, x(t, u(\cdot))) \leqq (v, y)$$

for all positive t and admissible controllers u.

We shall show that, under the conditions of the theorem statement, an upper bound on this inner product is impossible. As before the proof will be divided into the three cases (3.3), (3.4) and (3.5).

Proof for (3.3). The inner product $(v, x(t, u(\cdot)))$ can be written as

$$\begin{aligned}(v, x(t, u(\cdot))) &= \int_0^t (v, e^{A(t-s)}Bu(s))\,ds \\ &= \int_0^t (v, e^{As}Bu(t-s))\,ds.\end{aligned}$$

If u is a constant control function, then

$$(v, x(t, u(\cdot))) = \int_0^t (v, e^{As}Bu)\, ds.$$

From case (3.3) of Theorem 1.4 the inner product $(v, e^{As}Bu)$ is given by

$$(v, e^{As}Bu) = t^{j_1} e^{\lambda_1 t}((z_1, Bu) + F(t)),$$

where $F(t)$ vanishes as t becomes infinite. From condition (1.6) there is a u satisfying $(z_1, Bu) = \gamma > 0$. Thus,

$$(v, x(t, u)) = \int_0^t s^{j_1} e^{\lambda_1 s}(\gamma + F(s))\, ds.$$

For $s > t_0$, $F(s) < \gamma/2$. Thus, for large t,

$$(v, x(t, u)) = \int_0^{t_0} s^{j_1} e^{\lambda_1 s}(\gamma + F(s))\, ds + \int_{t_0}^t s^{j_1} e^{\lambda_1 s}(\gamma + F(s))\, ds$$

$$> K + \frac{\gamma}{2}\int_{t_0}^t s^{j_1} e^{\lambda_1 s}\, ds.$$

From this last inequality it is clear that the inner product $(v, x(t, u(\cdot)))$ is not bounded above for the case (3.3) since $\lambda_1 > 0$, and thus the last integral diverges.

Proof for (3.4). For this case we have

$$(v, e^{At}Bu_0) = t^{r_1} e^{\rho_1 t}(a(t) + H(t)),$$

where the vector u_0 can be chosen so that $a(t)$ is not identically zero. Recall that $a(t)$ is almost periodic with mean value 0 and that $H(t)$ vanishes as t becomes infinite. Apply Lemma 2.8 to the function a to get the numbers L and q and the set of open intervals of length L on which $a(t) > q$. Let T be a large positive number and define $u(s) = u_0$ if $T - s \in I_j$ for some j and $u(s)$ to be in the kernel of B if $T - s$ is not in any I_j. Let $I_j = (t_{2j-1}, t_{2j})$, and recall that $0 \leqq t_j < t_{j+1}$ and that $t_{2j} - t_{2j-1} = L$ for all $j = 1, \cdots, \infty$. Thus, the inner product

$$(v, x(t_{2l}, u(\cdot))) = \int_0^{t_{2l}} (v, e^{As}Bu(t_{2l} - s))\, ds$$

$$= \sum_{j=1}^{l} \int_{t_{2j-1}}^{t_{2j}} (v, e^{As}Bu_0)\, ds$$

$$= \sum_{j=1}^{l} \int_{t_{2j-1}}^{t_{2j}} t^{r_1} e^{\rho_1 t}(a(t) + G(t))\, dt.$$

For $t > t'$, $|G(t)| < q/2$, and thus, for $t_{2j-1} > t'$,

$$\int_{t_{2j-1}}^{t_{2j}} t^{r_1} e^{\rho_1 t}(a(t) + G(t))\, dt > \frac{q}{2} t_{2j-1}^{r_1} e^{\rho_1 t_{2j-1}} L.$$

Since t_{2j-1} goes to infinity as j goes to infinity and since ρ_1 is positive, these integrals become arbitrarily large. Therefore, the scalar products of the form $(v, x(t, u(\cdot)))$ are not bounded above.

Proof for (3.5). In this case the scalar product

$$(v, e^{At}Bu) = t^{j_1} e^{\lambda_1 t}((z_1, Bu) + a(t) + J(t)).$$

By hypothesis we can choose a vector u satisfying $(z_1, Bu) = \gamma > 0$. If the function a is identically zero, the proof of (3.5) is the same as for (3.3). If $a(t)$ is not identically zero, we can modify the proof of (3.4) by noting that

$$\int_{t_{2j-1}}^{t_{2j}} (v, e^{At}Bu)\,dt = \int_{t_{2j-1}}^{t_{2j}} t^{j_1} e^{\lambda_1 t}(\gamma + a(t) + J(t))\,dt$$

$$> \left(\gamma + \frac{q}{2}\right) t_{2j-1}^{j_1} e^{\lambda_1 t_{2j-1}}$$

for t_{2j-1} sufficiently large. Thus the inner products $(v, x(t, u(\cdot)))$ are not bounded above.

Thus, in all three cases, the conditions of the theorem are sufficient to ensure that the origin can be steered to any point in R^n by uniformly bounded controllers.

COROLLARY 3.6. *Consider the system* (1.1) *with bounded restraint set* Ω *satisfying* (1.3) *and* (1.4). *Necessary and sufficient conditions that any point can be steered to the origin in finite time are* (1.5), (1.6) *and that no eigenvalue of* A *has a positive real part.*

Proof. This result follows immediately from Theorem 3.5 if we reverse the sense of time in system (1.1). This time reversal gives the system (1.2), and we note that if no eigenvalue of A has positive real part, then no eigenvalue of $-A$ has negative real part. Thus, since the above conditions are necessary and sufficient to steer solutions of (1.2) to any point in R^n, it follows from Remark 1.1 that the above conditions are necessary and sufficient to steer any point in R^n to the origin.

COROLLARY 3.7. *The system* (1.1) *with bounded restraint set* Ω *satisfying* (1.3) *and* (1.4) *is completely controllable if and only if* (1.5) *and* (1.6) *are satisfied and all eigenvalues of* A *have zero real part.*

Proof. The proof is the same as that in [4, Theorem 5.4] since the proof requires only that the system be globally null-controllable and that the reachable sets be all of R^n. The results of Theorem 3.5 and Corollary 3.6 are necessary and sufficient to satisfy both of these requirements.

Certain results in linear control theory have the requirement that the origin be interior to the control region when all that is really required is controllability. The following result is an example of this [2, p. 128].

THEOREM 3.8. *Consider the autonomous linear control system* (1.1) *in* R^n *with compact restraint set* $\Omega \subset R^m$, *initial state* x_0, *and the origin as fixed target in* R^n. *Assume that* (a) $u = 0$ *lies in the interior of* Ω, (b) $C(A, B)$ *has rank* n, *and* (c) *each eigenvalue* λ *of* A *satisfies* $\operatorname{Re}\lambda < 0$. *Then there exists a minimal time-optimal controller* $u^*(t)$ *on* $0 \leqq t \leqq t^*$ *steering* x_0 *to the origin.*

Using the result of Corollary 3.6, we can extend this result to the following corollary.

COROLLARY 3.9. *Consider the system* (1.1) *in* R^n *with compact restraint set* $\Omega \subset R^m$, *initial state* x_0, *and the origin as a fixed target in* R^n. *Assume* (1.5), (1.6), *and that each eigenvalue of* A *satisfies* $\operatorname{Re}\lambda \leqq 0$. *Then there exists a minimal time-optimal controller* $u^*(t)$ *on* $0 \leqq t \leqq t^*$ *steering* x_0 *to the origin.*

Proof. The proof is exactly the same as that in [2, p. 128] since this proof requires only global null-controllability in R^n.

As a final corollary of Theorem 1.4 we obtain an application to nonlinear control systems. In reference [6] the following result appears.

THEOREM 3.10. *Consider the nonlinear control system*

$$\dot{x} = f(x, u) \tag{3.7}$$

with control restraint Ω *and* $f(0, 0) = 0$. *If the reachable set for the associated linear system*

$$\dot{y} = Ay + v(t),$$

where $A = f_x(0, 0)$ *and* $v(t) \in K(f(0, \Omega))$ *(the smallest closed cone containing* $f(0, \Omega)$ *with vertex at the origin) contains an open set about the origin, then the reachable set for the nonlinear system contains an open set about the origin.*

Using this theorem we can obtain the following corollary.

COROLLARY 3.11. *Consider the nonlinear control system* (3.7), *and assume that* $K(f(0, \Omega))$ *has nonempty interior in some subspace* S *of* R^n. *Let* E *denote the canonical injection from* S *considered as a vector space to* R^n, *and assume that the associated linear system*

$$\dot{x} = Ax + Ew, \qquad w \in S,$$

satisfies (1.5) *and* (1.6) *when* w *is constrained to the subset of* $f(0, \Omega)$ *contained in* S. *Then the reachable set for* (3.7) *contains an open neighborhood of the origin.*

Proof. The result is immediate by applying Theorem 1.4 to Theorem 3.11.

REFERENCES

[1] N. JACOBSON, *Lectures in Abstract Algebra*, vol. II, Van Nostrand, Princeton, N. J., 1953.

[2] E. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[3] S. SAPERSTONE, *Global controllability of linear systems using positive controls*, this Journal, to appear.

[4] S. SAPERSTONE AND J. YORKE, *Controllability of linear oscillatory systems using positive controls*, this Journal, 9 (1971), pp. 253–262.

[5] J. STIGLITZ, ed., *The Collected Scientific Papers of Paul A. Samuelson*, M.I.T. Press, Cambridge, Mass., 1966.

[6] J. YORKE, *The maximum principle and controllability of nonlinear equations*, this Journal, pp. 336–340.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

CERTAIN CONTROLLABILITY PROPERTIES OF ANALYTIC CONTROL SYSTEMS*

VELIMIR JURDJEVIC†

Abstract. In this article we consider analytic control systems defined on an analytic manifold M. In particular, we study the relationship between the accessibility property, i.e., attainable sets having nonempty interior, and controllability. This study is restricted to analytic control systems because they admit a very simple algebraic criterion for determining the accessibility property; this criterion is a generalization of the well-known rank condition for linear systems.

The main difficulty in extending the accessibility property to yield controllability lies in the fact that the time coordinate is nondecreasing. Beside systems of the form $F(x, u) = -F(x, -u)$, right invariant control systems defined on a compact Lie group are controllable whenever they have the accessibility property. It seems reasonable that a similar result should hold for arbitrary compact manifolds.

Introduction. In this article we shall consider the relationship between the accessibility property (i.e., attainable sets from each point having a nonempty interior) and controllability. In particular, we single out some systems in which the above notions are equivalent. This study will be restricted to systems of the form

$$dx/dt = F(x, u),$$

where, for each u, $F(\cdot, u)$ is an analytic function of the state variable x. We treat such systems because they have a very simple algebraic criterion for determining whether or not they have the accessibility property (cf. Sussmann and Jurdjevic [7]).

The main difficulty in extending the accessibility property to global results lies in the fact that the time coordinate is nondecreasing. In the case of symmetric systems, i.e., systems of the form $F(x, u) = -F(x, -u)$ this difficulty is removed because the points attainable in a negative time by the control u can be attained in positive time by $-u$. When the system is asymmetric the situation is not so simple. In fact, the linear systems offer a rare example of asymmetric systems in $\mathbb{R}^n$ with the accessibility property being equivalent to controllability.

Motivated by certain considerations in control of the rotations of a rigid body (Brockett [1]), we study the right-invariant control systems on a Lie group G. We demonstrate that when G is a compact connected Lie group, then the accessibility property and controllability are equivalent. Influenced by this result, we then raise similar questions concerning arbitrary compact manifolds.

1. Statement of the problem and the basic definition. We consider the state space M to be an n-dimensional real analytic manifold. M will always be assumed to be paracompact and connected. This condition is needed so that certain "obvious" results hold true. For instance, it in particular guarantees that all open submanifolds of M must be n-dimensional. We assume that the elements of the class of admissible controls $\mathscr{U}$ are defined on $[0, \infty)$ and have values in Ω, $\Omega \subset \mathbb{R}^m$,

* Received by the editors September 17, 1971, and in revised form November 1, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Department of Mathematics, University of Toronto, Toronto, Ontario, Canada.

$m \geqq 1$. The essential property of $\mathcal{U}$ is that it contains "sufficiently many" elements, and for that reason we assume that $\mathcal{U}$ contains all the piecewise constant Ω-valued functions defined on $[0, \infty)$. We shall assume that the control system is described by the following evolution equation:

$$dx(t)/dt = F(x(t), u(t)), \qquad u \in \mathcal{U}. \tag{1}$$

In addition, we shall assume that Ω, $\mathcal{U}$ and F satisfy further conditions which ensure that, corresponding to each $x \in M$, $u \in \mathcal{U}$, (1) admits a unique solution defined everywhere on $[0, \infty)$ and which depends continuously on all the parameters (here it is understood that $\mathcal{U}$ is endowed with an appropriate topology so that continuity makes sense). Since the precise statement of the previous conditions is lengthy and somewhat technical, it will be omitted. For the details the reader may consult [7]. F will be assumed to satisfy an additional property which is absolutely essential for the subsequent theory: we assume that for each $\omega \in \Omega$, $F(\cdot, \omega)$ is an analytic vector field on M. We shall also assume, for convenience, that each $F(\cdot, \omega)$ is complete, i.e., that the one-parameter group induced by $F(\cdot, \omega)$ is global. We term the above systems analytic control systems. Corresponding to each $x \in M$, $u \in \mathcal{U}$ we shall denote the solution of (1) by $\pi(x, u, \cdot)$. We shall say that $y \in M$ is attainable from $x \in M$ at time t, $t \geqq 0$, if there exists $u \in \mathcal{U}$ such that $\pi(x, u, t) = y$. y is attainable from x if there exists a time $t > 0$ such that y is attainable from x at t. The set of all elements attainable from x at time t will be denoted by $A(x, t)$, and the set of all elements attainable from x by $A(x)$. Equivalently, for each $x \in M$, $A(x) = \bigcup_{t \geqq 0} A(x, t)$. We shall say that the control system has the accessibility property from x if $A(x)$ has a nonempty interior in M. Analogously, if for some $t > 0$, $A(x, t)$ has a nonempty interior in M, then the system has the strong accessibility property from x. If the above properties hold for each $x \in M$, then the system is said to have the accessibility property (respectively, the strong accessibility property). The system is controllable from $x \in M$ at time $t > 0$ if $A(x, t) = M$, and analogously, it is controllable from x if $A(x) = M$. Finally, we say that the system is controllable if $A(x) = M$ for each $x \in M$.

2. Algebraic criteria of controllability. In this section we shall present certain algebraic criteria for $A(x)$ and $A(x, t)$ to have a nonempty interior.

Since these results were developed in full detail in [7], I shall sketch here only the main ideas.

Let $V(M)$ be the set of all analytic vector fields on M. We shall regard $V(M)$ as an algebra with the obvious operation of addition and the Lie product $[\cdot, \cdot]$ defined by $[X, Y] = X \cdot Y - Y \cdot X$ for X, Y in $V(M)$.

Let $D = \{F(\cdot, \omega) : \omega \in \Omega\}$. By our assumption D is a subset of $V(M)$. We shall denote by $I(D)$ the smallest subalgebra of $V(M)$ which contains D. Thus $I(D)$ contains D and all the linear combinations of the Lie products of all orders of elements of D. If $I(D)(x) = \{X(x) : X \in I(D)\}$, then we have that for each $x \in M$, $I(D)(x)$ is a linear subspace of the tangent space $T_x(M)$ at x; the dimension of $I(D)(x)$ may depend on x, and in that sense $I(D)$ is not a distribution on M in the terminology of Chevalley [2]. We are now going to use a modification of Chow's theorem [3] to obtain a characterization of the accessibility property solely in terms of dimensions of $I(D)(x)$.

Let $x \in M$ be arbitrary but fixed. If $X \in V(M)$, let X_t be the one-parameter transformation group generated by X; i.e., $t \to X_t(y)$ is an integral curve of X which passes through x at time $t = 0$. We define $\hat{A}(x)$ as follows:

$$\hat{A}(x) = \{X_{t_k}^k \cdot X_{t_{k-1}}^{k-1} \cdots X_{t_1}^1(x) : k > 0, \{X_1, \cdots, X_k\} \subset D, (t_1, \cdots, t_k) \in \mathbb{R}^k\}.$$

Obviously, $A(x) \cap \hat{A}(x)$ consists of all the elements

$$X_{t_k}^k \cdots X_{t_1}^1(x) \quad \text{with} \quad t_1 \geqq 0, t_2 \geqq 0, \cdots, t_k \geqq 0.$$

Now Chow's theorem asserts that $\hat{A}(x)$ is a submanifold of M which contains x, and whose tangent space at every point y is $I(D)(y)$. (Actually, Chow's theorem was stated only for the case when $\dim I(D)(x) = I(D)(y)$ for all x, y in M. Its modification to the present setting is provided by Lobry [6].) Therefore, $\hat{A}(x)$ has a nonempty interior in M if and only if $\dim I(D)(x) = \dim M$. Since we want an analogous condition stated in terms of $A(x)$ we proceed as follows: If $A(x)$ has a nonempty interior in M, then since $A(x) \subset \hat{A}(x)$ (and this follows almost immediately from the connectedness of integral curves of D), we have that $\hat{A}(x)$ has a nonempty interior in M. Hence, $\dim I(D)(x) = \dim M$. Conversely, if $\dim I(D)x = \dim M$, then $\hat{A}(x)$ has a nonempty interior in M. This implies that there exist $X_1, \cdots, X_k$ in D such that the image of the map F given by $(t_1, \cdots, t_k) \to X_{t_k}^k \cdot X_{t_{k-1}}^{k-1} \cdots X_{t_1}^1(x)$ has a nonempty interior. By Sard's theorem the set of points where the differential of F is of rank $< n$ must be of measure zero. Therefore, for some $\mathbf{t} \in \mathbb{R}^k$ the differential $dF_{\mathbf{t}}$ is of rank n. By analyticity of F, rank $dF_{\mathbf{t}} = n$ for an open and dense subset of $\mathbb{R}^k$. This shows that for some $t_1 > 0, t_2 > 0, \cdots, t_k > 0$, rank $dF_{\mathbf{t}} = \dim M$, which implies, by the implicit function theorem, that $A(x)$ has a nonempty interior in M. Thus, we have proved the following proposition.

PROPOSITION 1. *An analytic control system has the accessibility property from x if and only if* $\dim I(D)(x) = \dim M$. *It has the accessibility property if and only if* $\dim I(D)x = \dim M$ *for each* $x \in M$.

Proposition 1 has its analogue in terms of the strong accessibility property. We shall only state this analogue because its verification is quite similar to the verification of Proposition 1. Let D' be the linear subspace of $V(M)$ generated by all the elements of the form $[X, Y]$ with X and Y in $I(D)$. Let

$$I_0(D) = \left\{ \sum_{i=1}^{k} \lambda_i X_i + Y : k \geqq 1, \sum_{i=1}^{k} \lambda_i = 0, \{X_1, \cdots, X_k\} \subset D \text{ and } Y \in D' \right\}.$$

It is readily verifiable that $I_0(D)$ is an ideal of $I(D)$. Furthermore, we have the following proposition.

PROPOSITION 2. *An analytic control system has the strong accessibility property from x if and only if* $\dim I_0(D)(x) = \dim M$. *It has the strong accessibility property if and only if* $\dim I_0(D)(x) = \dim M$ *for each* $x \in M$.

Incidently, the preceding criterion implies that if $A(x, t)$ has a nonempty interior for some $t > 0$, then it must have a nonempty interior for all $t > 0$.

As a simple illustration of the above propositions we offer the following example.

Example 3. Let $M = \mathbb{R}^2$, $\Omega = \{0, 1\}$, and let $\mathcal{U}$ consist of all the piecewise constant functions on $[0, \infty)$ having values in Ω. Let $F(x, u) = A(u)x$, where $A(u)$

is a 2×2 matrix given by $A(u) = \begin{pmatrix} 1 & 0 \\ 1-u & 1 \end{pmatrix}$. Then $D = \{V_1, V_2\}$, where $V_1 x = x$, and $V_2 x = \begin{pmatrix} x_1 \\ x_1 + x_2 \end{pmatrix}$. Since V_1 and V_2 commute, $I(D) = \{\lambda V_1 + \mu V_2 : \lambda, \mu \text{ real}\}$, and therefore, $I_0(D) = \{\lambda(V_1 - V_2) : \lambda \text{ real}\}$.

By a simple computation we get that

$$\dim I(D)(x) = \begin{cases} 0 & \text{if and only if } x = 0, \\ 1 & \text{if } x_1 = 0 \text{ and } x_2 \neq 0, \\ 2 & \text{otherwise.} \end{cases}$$

Similarly,

$$\dim I_0(D)(x) = \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{otherwise.} \end{cases}$$

Thus, the system has no accessibility property, although for each $x \in \mathbb{R}^2$ with $x_1 \neq 0$, $A(x)$ has a nonempty interior. Furthermore, there are no points in $\mathbb{R}^2$ from which the system has the strong accessibility property.

Indeed, it is very simple to compute the attainable sets and to show that the attainable sets which originate on the x_2-axis remain there, and that the attainable sets which do not originate on the x_2-axis have a nonempty interior.

3. Controllability. In this section we shall demonstrate how the accessibility criteria of the previous section can be utilized to obtain global controllability results. As an introduction we start with the following well-known example.

Example 4 (The linear system). Let $M = \mathbb{R}^n$, $n \geqq 1$, $\Omega = \mathbb{R}^m$ and let $\mathcal{U}$ be the class of all piecewise constant functions defined on $[0, \infty)$ with values in Ω. Let $F(x, w) = Ax + Bw$ for all $x \in \mathbb{R}^n$, $w \in \mathbb{R}^m$, where A and B are matrices of dimensions $n \times n$ and $n \times m$ respectively.

If $\pm w_i$ is the vector in $\mathbb{R}^m$ whose ith component is ± 1 and whose all other components are zero, we get that $F(\cdot, \pm w_i) = A(\cdot) \pm b_i$, where b_i is the ith column of the matrix B.

Thus $\{A(\cdot) \pm b_i : i = 1, \cdots, m\} \subset D$, and by a direct computation it follows that $I(D)(x)$ contains vectors $Ax \pm b_i, \pm Ab_i, \cdots, \pm A^{n-1} b_i$ for $i = 1, 2, \cdots, m$. In particular, when $x = 0$, we have that

$$I(D)(0) = I_0(D)(0)(b_i = \tfrac{1}{2} b_i + (-\tfrac{1}{2})(-b_i) \text{ for } i = 1, 2, \cdots, m).$$

From the above, it follows immediately that $A(0)$ has a nonempty interior if and only if $A(0, t)$ has a nonempty interior for some $t > 0$ (which, as remarked in the previous section, implies that $A(x, t)$ must have a nonempty interior for all $t > 0$). Since $\mathcal{U}$ is "unrestricted" and is closed under concatenations, we have that $A(0)$ and $A(0, t)$ (for each $t > 0$) are linear subspaces of $\mathbb{R}^n$. This fact, together with the above, implies that $A(0, t)$ has a nonempty interior if and only if $A(0, t) = \mathbb{R}^n$ for each $t > 0$. Since, obviously, $\dim I(D)(0) = n$ if and only if $\dim I(D)(x) = n$ for each $x \in M$, we have that the accessibility property is equivalent to controllability.

In general, this situation is false. $\left(\text{Take, for instance, } D = \left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}.\right)$

Before we consider the next class of control systems motivated by the control of the rotations of a rigid body (cf. [1] and [5]), we state the following obvious interpretation of Chow's theorem.

PROPOSITION 5. *Let Ω be symmetric about the origin (i.e., $\omega \in \Omega$ implies that $-\omega \in \Omega$), and let an analytic control system be described by $\dot{x} = F(x, u)$, where $F(x, \omega) = -F(x, -\omega)$ for each $x \in M$, $\omega \in \Omega$. Then the accessibility property implies controllability.*

The proof essentially consists of observing that in this case $\hat{A}(x) = A(x)$ for each $x \in M$; because, if $y = X_t(x)$ for some $X \in D$, then $-X \in D$ and $y = -X_{-t}(x)$; i.e., the set of points attainable by an integral curve $t \to X_t(x)$ for $t < 0$ is the same as the set of points attainable by the curve $t \to -X_t(x)$ for $t > 0$. And since $\hat{A}(x)$ is the maximal integral manifold of M of dimension $=$ dim M, and M is connected, we have that $\hat{A}(x) = M$; and hence, $A(x) = M$.

The above case was termed by Lobry as a "symmetrical control problem" [6]. We now turn our attention to certain asymmetric systems which are motivated by control of rotational dynamics.

We assume that M is a connected Lie group G, and that our control system is described by the following equation:

$$\frac{dx(t)}{dt} = X(u(t))(x(t)), \qquad u \in \mathcal{U}, \tag{2}$$

where for each $\omega \in \Omega$, $X(\omega)$ is a right-invariant vector field on G. Without loss of generality, we shall assume that $\mathcal{U}$ consists of only piecewise constant functions on $[0, \infty)$.

We shall term the above systems as the right-invariant control systems. From the right-invariance of our control system, it immediately follows that for any $x \in G, t > 0, A(x, t) = A(e, t)x$, where e is the identity of G. (Here, $A \cdot x = \{ax : a \in A\}$.)

Therefore, a right-invariant system is controllable if and only if $A(e) = G$. Thus we restrict our attention to the sets attainable from the identity. Let x and y belong to $A(e)$ with $x = \pi(e, u, t)$, $y = \pi(e, v, s)$ for some u, v in $\mathcal{U}$ and $t > 0, s > 0$.

Let

$$w(\tau) = \begin{cases} u(\tau), & 0 \leqq \tau \leqq t, \\ v(\tau - t), & t < \tau. \end{cases}$$

Then, $w \in \mathcal{U}$ and $\pi(e, w, t + s) = yx$. Therefore, $A(e)$ is a semigroup in G. In general, $A(e)$ is not a group (cf. [5]). However, if G is a compact Lie group, then we have the following result.

PROPOSITION 6. *If a right-invariant control system defined on a connected compact Lie group G has the accessibility property, then it is controllable there.*

The proof consists of several steps:

First we show that the closure of $A(e)$ is a subgroup of G. Let $x \in \overline{A(e)}$. Since $A(e)$ is a semigroup, so is its closure $\overline{A(e)}$. Therefore, for each positive integer n, $x^n \in \overline{A(e)}$. By compactness of G, it follows that there is a subsequence $\{n_k\}$ such that $\{x^{n_k}\}$ converges. Without loss of generality, assume that $n_{k+1} > n_k$. We have that

$x^{-1} = \lim x^{n_{k+1}-n_k-1}$. Since $n_{k+1} - n_k - 1 \geqq 0$, we have that $x^{n_{k+1}-n_k-1} \in \overline{A(e)}$, and hence $x^{-1} \in \overline{A(e)}$. Therefore, $\overline{A(e)}$ is a group.

By the accessibility property $\overline{A(e)}$ has a nonempty interior in G, and since G is connected, $\overline{A(e)} = G$. Thus $A(e)$ is dense in G.

Finally we show that the above imply that $A(e) = G$. Let $x \in \operatorname{int} A(e)$, and let U be an open neighborhood of x which is contained in $A(e)$. Let $V = \{y^{-1} : y \in U\}$. V is open in G, and since $A(e)$ is dense in G, $A(e) \cap V \neq \varnothing$. Let $y \in A(e) \cap V$. Then, Uy is an open neighborhood of the identity which is contained in a semigroup $A(e)$; hence, $A(e)$ is a group which, in addition, has a nonempty interior in G. Therefore, $A(e) = G$, and our proof is now complete.

In this situation, the strong accessibility property implies even more.

PROPOSITION 7. *If a right-invariant control system on a compact group G has the strong accessibility property, then there exists a time $T > 0$ such that $A(e, T) = G$.*

Proof. By the previous proposition, $A(e) = G$. For each $t > 0$, let $V(t) = \operatorname{int} A(e, t)$. If $x \in V(t)$, let $x^{-1} \in A(e, s)$ for some $s > 0$. Then, $e \in V(t + s)$. Thus, for $\tau = t + s$, $A(e, \tau)$ contains a neighborhood U of the identity. Furthermore, we have that for each integer $n > 0$, $U^n \subset A(e, n\tau)$. Since,

$$G = \bigcup_{n>0} U^n \subset \bigcup_{n>0} A(e, n\tau) \subset G,$$

it follows by compactness of G that for some integer $N > 0$, $G = \bigcup_{n<N} A(e, N\tau)$. Let $T = N\tau$. Since the sets $A(e, n\tau)$ are nondecreasing as n increases, we have that $G = A(e, T)$, and our proof is complete.

The next example shows that in the absence of compactness a right-invariant control system may have the accessibility property without being controllable. This example is due to H. Sussmann (cf. [5]).

Example 8. Let $G = SL(2, \mathbb{R})$, the group of all 2×2 real matrices whose determinant is 1. Let

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

and let $\mathcal{U}$ consist of all piecewise constant real functions on $[0, \infty)$. Consider the following right-invariant control system on G:

$$dx(t)/dt = (A + uB)x(t).$$

Since the trace of $A + u(t)B$ is zero for all $t > 0$, the problem is well-posed.

If x_{ij} are the components of an arbitrary solution which originates at the identity, then

$$\dot{x}_{11} = x_{11} + ux_{21} \quad \text{and} \quad \dot{x}_{21} = ux_{11} - x_{12}.$$

Multiplying the first equation by x_{11}, the second by x_{21} and subtracting we get

$$\frac{1}{2}\frac{d}{dt}(x_{11}^2 - x_{21}^2) = x_{11}^2 + x_{21}^2.$$

Thus, $x_{11}^2 - x_{21}^2$ is a nondecreasing function of time. Since $(x_{11}^2 - x_{21}^2)(0) = 1$, it follows that if $x \in A(e)$, then $x_{11}^2 \geqq 1 + x_{21}^2$. Therefore, the system is not controllable.

4. Remarks. (i) In both the linear case and the right-invariant case the accessibility from a particular point is equivalent to accessibility, which in turn, is equivalent to controllability. This deduction of global results from the knowledge of behavior of the system at a single point will in general be a very difficult problem because such results would answer the problem of the singularity of vector fields which, to quote Hermann [4], "would include the problem of singularity of mappings which is already enormously difficult."

(ii) The preceding considerations of right-invariant control systems motivate the following question:

Are there analytic control systems defined on a compact manifold which have the accessibility property but which are not controllable?

REFERENCES

[1] R. W. BROCKETT, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–284.

[2] C. CHEVALLEY, *Theory of Lie Groups*, Princeton University Press, Princeton, N.J., 1946.

[3] W. L. CHOW, *Uber Systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann., 117 (1939), pp. 98–105.

[4] R. HERMANN, *E. Cartan's geometric theory of partial differential equations*, Advances in Math., 1 (1965), pp. 265–315.

[5] V. JURDJEVIC AND H. SUSSMANN, *Control systems on Lie groups*, J. Differential Equations, to appear.

[6] C. LOBRY, *Contrôlabilité des systemes non-linéaires*, this Journal, 8 (1970), pp. 573–604.

[7] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, to appear.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

ON THE CONTROL OF DISCRETE-TIME DISTRIBUTED PARAMETER SYSTEMS*

KWANG YUN LEE†, SHUI-NEE CHOW‡ AND ROBERT O. BARR†

Abstract. A system described by a difference equation with a semigroup of operators is considered. The existence and uniqueness of an optimal control for the minimization of a quadratic cost functional over a finite interval is proved. The necessary condition for optimal control is derived and the optimal control is given by a linear state feedback law. The feedback operator is shown to be bounded, positive and self-adjoint, and the optimal cost is expressed in terms of the feedback operator. The control on the infinite interval and the behavior as $n \to \infty$ are considered.

1. Introduction. Difference equations arise and are of utmost importance in the fields of, for example, numerical analysis and sampled-data control systems. Distributed parameter or infinite-dimensional systems are of particular interest, and a certain class of these has been studied in [1].

The purpose of this paper is to investigate the discrete-time distributed parameter system which has a semigroup of operators as the system operator and a quadratic criterion for the cost functional. The finite-dimensional control problem has been extensively investigated and can be found, for example, in [2], [3], [4], the results of which are similar to those obtained in this paper. The theory developed here parallels, to some extent, that given in [5] for the continuous-time problem.

2. Statement of the problem. Let H be a Hilbert space, and $x_i \in H$ be the state of the system at time i, where $i \in \sigma = \{0, 1, 2, \cdots, N\}$. Let U be a Hilbert space of controls. Let $\Phi : H \to H$ and $D : U \to H$ be linear and continuous.

The control system is given by

$$x_{i+1} = \Phi x_i + Du_i, \quad x_0 \in H, \tag{2.1}$$

and the quadratic cost functional is given by

$$J(u) = \sum_{i=0}^{N-1} [\langle x_i, Qx_i \rangle + \langle u_i, Ru_i \rangle], \tag{2.2}$$

where $Q : H \to H$ is a bounded self-adjoint positive semidefinite operator and $R : U \to U$ is a bounded self-adjoint positive definite operator; i.e., for some positive numbers b_0, b_1, and b_2, $0 \leqq \langle x_i, Qx_i \rangle \leqq b_0 \|x_i\|^2$ for all $x_i \in H$ and $b_1 \|u_i\|^2 \leqq \langle u_i, Ru_i \rangle \leqq b_2 \|u_i\|^2$ for all $u_i \in U$, where $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the inner product and norm on H, respectively (or on U, which can be distinguished in the context).

* Received by the editors September 13, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Department of Electrical Engineering and Systems Science, Michigan State University, East Lansing, Michigan 48823. The work of these authors was supported in part by the National Science Foundation under Grant GI-20.

‡ Department of Mathematics, Michigan State University, East Lansing, Michigan 48823. The work of this author was supported in part by the National Science Foundation under Grant GU-2648.

Now we state the optimal control problem.

Problem 2.1. The optimal control problem for the system (2.1) is to determine the control sequence $\{u_i^*\}_{i\in\sigma}$ with $u_i^* \in U$ such that the functional $J(u)$ in (2.2) is minimized at $\{u_i^*\}_{i\in\sigma}$ for all sequences $\{u_i\}_{i\in\sigma}$ with $u_i \in U$.

Let $l^2(\sigma; H)$ be the family of all sequences $\{x_i\}_{i\in\sigma}$ with $x_i \in H$.

Remark 2.1 [6, p. 257]. $l^2(\sigma; H)$ is a Hilbert space with the usual addition and scalar multiplication, and with an inner product defined by, for $x = \{x_i\}$ and $y = \{y_i\}$ in $l^2(\sigma; H)$,

$$\langle x, y\rangle_{l^2(\sigma;H)} = \sum_{i=0}^{N-1} \langle x_i, y_i\rangle_H. \tag{2.3}$$

Similarly, we may define a Hilbert space $l^2(\sigma; U)$ with inner product analogous to (2.3).

Remark 2.2. The assumption that Φ is a bounded operator is not so restrictive even in the distributed parameter system. For example, consider a sampled-data control problem; i.e., sampling the continuous-time parabolic system given by

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t). \tag{2.4}$$

Here the elliptic partial differential operator A is the infinitesimal generator of a strongly continuous semigroup of operators $\Phi(t)$ and the domain of A is dense in H. B is a bounded operator from U into H. The system (2.4) can be transformed to the equivalent discrete-time system

$$x_{i+1} = \Phi x_i + Du_i, \tag{2.5}$$

where $x_i = x(t_i)$, $\Phi = \Phi(\delta)$, δ is a sampling period, and D is a bounded operator from U into H such that

$$Du_i = \int_0^\delta \Phi(\delta - t)Bu_i\, dt. \tag{2.6}$$

Clearly, we have the system operator Φ as a bounded operator. The details will be found in a separate paper [8].

In order to prove the existence and uniqueness of Problem 2.1, we require that the solution of difference equation (2.1) depend continuously on the control. Hence we state the following lemma which will be found in [8].

LEMMA 2.1. *The mapping $u \to x$ of $l^2(\sigma, U)$ into $l^2(\sigma; H)$ defined by the difference equation* (2.1) *is continuous.*

3. Solution of the problem. Lions [5] proved a general existence and uniqueness theorem for controls minimizing a certain cost functional. He also showed that this theorem covers the existence and uniqueness of optimal controls for the continuous parabolic control problem. In this section the discrete-time problem, Problem 2.1, will be shown to fall into Lions' framework in the function spaces $l^2(\sigma; H)$ and $l^2(\sigma; U)$.

THEOREM 3.1. *The discrete-time problem as specified in Problem* 2.1 *has a unique solution $u^* \in l^2(\sigma; U)$.*

To prove this we need the following definition and lemmas which can be found in [5]. Let V be a Hilbert space.

DEFINITION 3.1. A continuous symmetric coercive bilinear form $\pi(u, v)$ is a continuous function in both arguments which maps $V \times V$ into the real numbers for which there exists a $c > 0$ such that

$$\pi(u, v) \geqq c\|u\|^2 \quad \text{for all } u \in V,$$

$$\pi(u, v) \equiv \pi(v, u) \quad \text{for all } u, v \in V.$$

Consider a functional

$$C(u) = \pi(u, u) - 2L(u), \qquad u \in V, \tag{3.1}$$

where L is a bounded linear functional defined on V.

LEMMA 3.1. *If $\pi(u, v)$ is a continuous symmetric coercive bilinear form, then there exists a unique $u^* \in V$ such that*

$$C(u^*) = \inf_{i \in V} C(u).$$

LEMMA 3.2. *If the hypotheses of Lemma* 3.1 *are satisfied, then the minimizing element $u^* \in V$ is characterized by*

$$\pi(u^*, v) = L(v) \quad \textit{for all } v \in V. \tag{3.2}$$

Now we are in the position to prove Theorem 3.1.

Proof of Theorem 3.1. We may write the cost functional (2.2) with inner products in the function spaces $l^2(\sigma; H)$ and $l^2(\sigma; U)$; that is,

$$J = \langle x, Qx \rangle_{l^2(\sigma;H)} + \langle u, Ru \rangle_{l^2(\sigma;U)}, \tag{3.3}$$

where Q and R are bounded operators on $l^2(\sigma; H)$ and $l^2(\sigma; U)$, respectively, with ranges in $l^2(\sigma; H)$ and $l^2(\sigma; U)$, respectively, such that

$$Qx = \{Qx_i\} \quad \text{and} \quad Ru = \{Ru_i\}, \qquad i \in \sigma.$$

Note that Q and R are self-adjoint and are positive semidefinite and positive definite operators, respectively.

Let x^u denote the solution of (2.1) with control $u \in l^2(\sigma; U)$. We define the bilinear form $\pi(u, v)$ on $l^2(\sigma; U) \times l^2(\sigma; U)$ to be

$$\pi(u, v) \equiv \langle x^v - x^0, Q(x^u - x^0) \rangle_{l^2(\sigma;H)} + \langle v, Ru \rangle_{l^2(\sigma;U)}, \tag{3.4}$$

and the linear functional $L(v)$ on $l^2(\sigma; U)$ to be

$$L(v) \equiv -\langle x^v - x^0, Qx^0 \rangle_{l^2(\sigma;H)}. \tag{3.5}$$

Then the cost functional J in (3.3) becomes

$$J(u) = \pi(u, u) - 2L(u) + \langle x^0, Qx^0 \rangle_{l^2(\sigma;H)}.$$

Since the last term is independent of u, minimizing the cost functional

$$J_1(u) = \pi(u, u) - 2L(u)$$

is equivalent to minimizing the original cost functional $J(u)$. The continuity of $\pi(u, v)$ follows from Lemma 2.1. The symmetricity of $\pi(u, v)$ can be shown by

using the fact that Q and R are self-adjoint. The coercivity of $\pi(u, v)$ follows directly from the fact that Q is positive semidefinite and R is positive definite. Since $\pi(u, v)$ satisfies the hypotheses of Lemma 3.1, there exists a unique $u^* \in l^2(\sigma; U)$ such that

$$J_1(u^*) = \inf_{u \in l^2(\sigma;U)} J_1(u).$$

Next we derive the necessary condition for optimality which is analogous to the result for finite-dimensional systems.

THEOREM 3.2. *If $u^* \in l^2(\sigma; U)$ is the optimal control for Problem 2.1 with optimal response $x^* \in l^2(\sigma; H)$, then there necessarily exists a unique adjoint state $p^* \in l^2(\sigma; H)$ such that*

$$u_i^* = -R^{-1}D^*p_{i+1}^*, \tag{3.6}$$

$$p_i^* = \Phi^*p_{i+1}^* + Qx_i^*; \qquad p_N^* = 0, \tag{3.7}$$

where R^{-1} is the inverse of R, and D^* and Φ^* are the adjoints of D and Φ, respectively.

Proof. We note x^* satisfies

$$x_{i+1}^* = \Phi x_i^* + Du_i^*; \qquad x_0^* = x_0. \tag{3.8}$$

From the proof of Theorem 3.1, $\pi(u, v)$ is a continuous symmetric coercive bilinear form. Hence, by Lemma 3.2, the optimal control must satisfy

$$\pi(u^*, v) = L(v) \quad \text{for all } v \in l^2(\sigma; U),$$

or equivalently.

$$\pi(u^*, v - u^*) = L(v - u^*) \quad \text{for all } v \in l^2(\sigma; U). \tag{3.9}$$

Further, let us introduce the adjoint equation

$$\begin{aligned} p_i &= \Phi^*p_{i+1} + Qx_i, \\ p_N &= 0. \end{aligned} \tag{3.10}$$

Here p_i is called the adjoint state and it should be noted that a unique solution $p \in l^2(\sigma; H)$ exists for (3.10). In fact, by changing i to $N - i$ and realizing Φ^* is a semigroup, we can have an explicit solution for p_i for all $i \in \sigma$. Let us denote p^u as the solution p due to the application of control $u \in l^2(\sigma; U)$. Forming the inner product on $l^2(\sigma; H)$ with $x^v - x^u$ we obtain

$$\begin{aligned} \langle p^u, x^v - x^u \rangle_{l^2(\sigma;H)} &= \sum_{i=0}^{N-1} \langle p_i^u, x_i^v - x_i^u \rangle \\ &= \sum_{i=0}^{N-1} \langle p_{i+1}^u, \Phi(x_i^v - x_i^u) \rangle + \langle Qx^u, x^v - x^u \rangle_{l^2(\sigma;H)}. \end{aligned} \tag{3.11}$$

Observing that $\langle p_i^u, x_i^v - x_i^u \rangle = 0$ for $i = 0$ and $i = N$, we see that the left-hand side of (3.11) becomes

$$\langle p^u, x^v - x^u \rangle_{l^2(\sigma;H)} = \sum_{i=0}^{N-1} \langle p_{i+1}^u, x_{i+1}^v - x_{i+1}^u \rangle. \tag{3.12}$$

Equating (3.11) and (3.12), using (2.1), and letting $u = u^*$ we obtain

$$\langle Qx^{u^*}, x^v - x^{u^*}\rangle_{l^2(\sigma;H)} = \sum_{i=0}^{N-1} \langle p_{i+1}^{u^*}, D(v_i - u_i^*)\rangle. \tag{3.13}$$

Now, from (3.9),

$$\begin{aligned} 0 &= \pi(u^*, v - u^*) - L(v - u^*) \\ &= \langle Q(x^{u^*} - x^0), x^v - x^{u^*}\rangle_{l^2(\sigma;H)} + \langle Ru^*, v - u^*\rangle_{l^2(\sigma;U)} \\ &\quad + \langle Qx^0, x^v - x^{u^*}\rangle_{l^2(\sigma;H)} \\ &= \langle Qx^{u^*}, x^v - x^{u^*}\rangle_{l^2(\sigma;H)} + \langle Ru^*, v - u^*\rangle_{l^2(\sigma;U)}. \end{aligned} \tag{3.14}$$

Combining (3.13) and (3.14), we obtain

$$\sum_{i=0}^{N-1} \langle p_{i+1}^{u^*}, D(v_i - u_i^*)\rangle + \langle Ru^*, v - u^*\rangle_{l^2(\sigma;U)} = 0,$$

or equivalently,

$$\sum_{i=0}^{N-1} [\langle D^* p_{i+1}^{u^*}, v_i - u_i^*\rangle + \langle Ru_i^*, v_i - u_i^*\rangle] = 0. \tag{3.15}$$

Since the equality (3.15) must hold for all $v \in l^2(\sigma; U)$, it must be true that

$$Ru_i^* = -D^* p_{i+1}^{u^*} \tag{3.16}$$

is satisfied by the optimal control u^*. Moreover, since R is positive definite it has an inverse and so (3.16) reduces to

$$u_i^* = -R^{-1} D^* p_{i+1}^{u^*}.$$

4. Decoupling and the Riccati operator difference equation. In this section we derive a feedback form of the optimal control given by (3.6). The feedback operator is shown to be bounded, self-adjoint and positive semidefinite, and the cost functional is expressed in terms of the feedback operator.

Consider the system of equations (3.7) and (3.8) with optimal control in (3.6) which can be expressed in the form:

$$\begin{aligned} x_{i+1} &= \Phi x_i - DR^{-1} D^* p_{i+1}, \\ p_i &= \Phi^* p_{i+1} + Qx_i, \\ x_s &= h \in H, \quad p_N = 0; \\ & \qquad i \in \{s, s+1, \cdots, N\}, \quad s \in \sigma. \end{aligned} \tag{4.1}$$

This system admits a unique solution pair $(x, p) \in l^2(s, N; H) \times l^2(s, N; H)$. This fact is easily seen if the cost functional J in (2.2) is defined on the interval $[S, N)$ instead of on $[0, N)$.

LEMMA 4.1. *The mapping* $h \to (x, p) =$ *solution of* (4.1) *is continuous from* H *into* $l^2(s, N; H) \times l^2(s, N; H)$.

Proof. Let us denote by $x^n(v)$ the state of the system given by

$$\begin{aligned} x_{i+1} &= \Phi x_i + Dv \quad \text{on } [s, N), \\ x_s &= h_n. \end{aligned} \tag{4.2}$$

For a fixed v, if $h_n \to h$,

$$x^n(v) \to x(v) \quad \text{in } l^2(s, N; H). \tag{4.3}$$

Let u^n (resp. u) be the optimal control for $J_s^{h_n}(v)$ (resp. $J_s^h(v)$). Hence, $J_s^{h_n}(u^n) = \inf J_s^{h_n}(v) \leqq J_s^{h_n}(u)$ and $J_s^{h_n}(u) \to J_s^h(u)$ from (4.3). Hence,

$$\limsup J_s^{h_n}(u^n) \leqq J_s^h(u) = \inf J_s^h(v). \tag{4.4}$$

But

$$J_s^{h_n}(u^n) \geqq C \sum_{i=s}^{N-1} \|u_i^n\|^2$$

which when combined with (4.4) shows that u^n belongs to a bounded subset of $l^2(s, N; U)$ as $h_n \to h$.

We then take a subsequence u^m such that

$$u^m \to w \quad \text{weakly in } l^2(s, N; U). \tag{4.5}$$

Therefore, $x^m(u^m) \to x(w)$ weakly in $l^2(s, N; H)$ and hence

$$\liminf J_s^{h_m}(u^m) \geqq J_s^h(w),$$

which when combined with (4.4) shows $J_s^h(w) \leqq J_s^h(u)$ and hence necessarily $w = u$. Therefore,

$$\begin{aligned} &u^n \to u \quad \text{weakly in } l^2(s, N; U), \\ &J_s^{h_n}(u^n) \to J_s^h(u), \end{aligned} \tag{4.6}$$

and

$$\begin{aligned} &x^n(u^n) \to x(u) \quad \text{weakly in } l^2(s, N; H), \\ &p^n(u^n) \to p(u) \quad \text{weakly in } l^2(s, N; H). \end{aligned}$$

This proves the continuity of the linear mapping $h \to (x, p)$ from H into $l^2(s, N; H) \times l^2(s, N; H)$. Furthermore, (4.6) implies that $u^n \to u$ strongly in $l^2(s, N; U)$ and hence the mapping $h \to (x, p)$ is in fact continuous in the strong topologies.

COROLLARY 4.1. *For $h \in H$, let (x, p) be the solution of* (4.1). *Then the mapping*

$$h \to p_s \tag{4.7}$$

is continuous from H into H.

Proof. The proof follows from the fact that the mapping defined by (4.7) is the composition of the mapping $h \to (x, p)$ and the mapping $(x, p) \to p_s$. But $(x, p) \in l^2(x, N; H) \times l^2(x, N; H)$ implies that for every $i \in [s, N)$, $\|x_i\|$ and $\|p_i\|$ are bounded. Hence we may take subsequences x_i^n, p_i^n such that

$$\begin{aligned} &x_i^n \to \bar{x}_i \quad \text{weakly in } H \text{ for all } i \in [s, N), \\ &p_i^n \to \bar{p}_i \quad \text{weakly in } H \text{ for all } i \in [s, N). \end{aligned}$$

The second equation of (4.1) becomes, in the limit,

$$\bar{p}_i = \Phi^* \bar{p}_{i+1} + Q\bar{x}_i \quad \text{on } [s, N);$$

hence, we may take $\bar{p}_i = p_i$ for all $i \in [s, N)$, and in particular, $\bar{p}_s = p_s$. Thus the mapping $(x, p) \to p_s$ is continuous from $l^2(s, N; H) \times l^2(s, N; H)$ into H.

THEOREM 4.1. *The optimal control $u^* \in l^2(\sigma; U)$ for Problem 2.1 is given by the feedback form*

$$u_i = -R^{-1}D^*K_{i+1}(I + DR^{-1}D^*K_{i+1})^{-1}\Phi x_i, \tag{4.8}$$

where K_i, $i \in \sigma$, is the solution of the Riccati-type operator difference equation

$$\begin{aligned} K_i &= \Phi^*K_{i+1}(I + DR^{-1}D^*K_{i+1})^{-1}\Phi + Q, \\ K_N &= 0. \end{aligned} \tag{4.9}$$

Proof. From the continuity of the transformation $h \to p_s$, p_s can be written uniquely in the form

$$p_s = K_s h. \tag{4.10}$$

Since s is arbitrary in σ and h is the evaluation of x_s, (4.10) implies that

$$p_i = K_i x_i \quad \text{for all } i \in \sigma, \tag{4.11}$$

where (x, p) is the solution pair of the system (4.1).

Using (4.11) we can rewrite the system (4.1) as

$$\begin{aligned} x_{i+1} &= \Phi x_i - DR^{-1}D^*K_{i+1}x_{i+1}, \\ p_i &= \Phi^*K_{i+1}x_{i+1} + Qx_i \\ &= K_i x_i, \qquad x_0 \in H; \quad p_N = 0. \end{aligned} \tag{4.12}$$

Rearranging the first equation in (4.12), we obtain

$$x_{i+1} = (I + DR^{-1}D^*K_{i+1})^{-1}\Phi x_i, \tag{4.13}$$

where the inverse is well-defined since K_{i+1} is positive semidefinite by the next theorem (Theorem 4.2). Substituting (4.13) into the second equation in (4.12) and rearranging terms, we have

$$[K_i - \Phi^*K_{i+1}(I + DR^{-1}D^*K_{i+1})^{-1}\Phi - Q]x_i = 0. \tag{4.14}$$

Since x_i is arbitrary in the sense that it depends on an arbitrary choice of x_0, it must be true that

$$\begin{aligned} K_i &= \Phi^*K_{i+1}(I + DR^{-1}D^*K_{i+1})^{-1}\Phi + Q, \\ K_N &= 0. \end{aligned}$$

Equation (4.8) follows from (3.6) with (4.11) and (4.13). This completes the proof.

Let $J_s^h(u)$ denote the cost (2.2) starting at s with initial state $h \in H$. We examine properties of K_s and express the optimal cost in terms of K_s.

THEOREM 4.2. *The operator $K_s \in \mathcal{L}(H, H)$ is self-adjoint, positive semidefinite, and bounded. Moreover, the optimal cost of system* (4.1) *starting at time s with initial state $h \in H$ is given by*

$$J_s^h(u^*) = \langle K_s h, h \rangle. \tag{4.15}$$

Proof. Proof of the self-adjointness. Let (x^1, p^1) and (x^2, p^2) be solution pairs corresponding to initial conditions h^1 and h^2, respectively, for system (4.1). Then

$$\begin{aligned} 0 &= \sum_{i=s}^{N-1} \langle p_i^1 - \Phi^* p_{i+1}^1 - Qx_i^1, x_i^2 \rangle \\ &= \sum_{i=s}^{N-1} \langle p_i^1, x_i^2 \rangle - \sum_{i=s}^{N-1} \langle p_{i+1}^1, \Phi x_i^2 \rangle - \sum_{i=s}^{N-1} \langle Qx_i^1, x_i^2 \rangle \\ &= \langle p_s^1, x_s^2 \rangle - \sum_{i=s}^{N-1} \langle p_{i+1}^1, DR^{-1}D^* p_{i+1}^2 \rangle - \sum_{i=s}^{N-1} \langle Qx_i^1, x_i^2 \rangle. \end{aligned}$$

Hence, by using (4.4),

$$\langle K_s h^1, h^2 \rangle = \sum_{i=s}^{N-1} \langle p_{i+1}^1, DR^{-1}D^* p_{i+1}^2 \rangle + \sum_{i=s}^{N-1} \langle Qx_i^1, x_i^2 \rangle. \tag{4.16}$$

Self-adjointness of K_s comes from those of R and Q.

Proof of (4.15) *and the positivity.* Let $u^* \in l^2(s, N; U)$ be the optimal control with cost $J_s^h(u^*)$. Optimality implies that u^* satisfies the necessary condition

$$Ru_i^* + D^* p_{i+1} = 0. \tag{4.17}$$

If $h^1 = h^2 = h$, the equality (4.16) becomes

$$\langle K_s h, h \rangle = \langle Qx, x \rangle_{l^2(s,N;H)} + \sum_{i=s}^{N-1} \langle p_{i+1}, DR^{-1}D^* p_{i+1} \rangle. \tag{4.18}$$

But, by virtue of (4.11),

$$\begin{aligned} \langle Ru^*, u^* \rangle_{l^2(s,N;U)} &= \sum_{i=s}^{N-1} \langle D^* p_{i+1}, R^{-1}D^* p_{i+1} \rangle \\ &= \sum_{i=s}^{N-1} \langle p_{i+1}, DR^{-1}D^* p_{i+1} \rangle. \end{aligned} \tag{4.19}$$

Combining (4.18) and (4.19), we obtain

$$\begin{aligned} \langle K_s h, h \rangle &= \langle Qx, x \rangle_{l^2(s,N;H)} + \langle Ru^*, u^* \rangle_{l^2(s,N;H)} \\ &= J_s^h(u^*) \geqq 0 \end{aligned} \tag{4.20}$$

which proves (4.15) and the positive semidefinite property of K_s.

Proof of the boundedness. Since the transformations $h \to x$ and $h \to p$ are continuous in the strong topology, we have $\|x\|_{l^2(s,N;H)} \leqq C_1\|h\|$ and $\|p\|_{l^2(s,N;H)} \leqq C_2\|h\|$. Since Q and R are bounded, we have

$$\langle Qx, x \rangle_{l^2(s,N;H)} \leqq M_1 C_1^2 \|h\|^2,$$

and

$$\sum_{i=s}^{N-1} \langle p_{i+1}, DR^{-1}D^* p_{i+1} \rangle \leqq M_2 C_2^2 \|h\|^2.$$

This implies that, by (4.18),

$$\begin{aligned} \langle K_s h, h\rangle &= J_s^h(u^*) \\ &\leqq (M_1 C_1^2 + M_2 C_2^2)\|h\|^2 = C\|h\|^2, \end{aligned} \tag{4.21}$$

which proves the boundedness of K_s. Hence the proof of the theorem is completed.

Remark 4.1. The existence and uniqueness of the solution of the Riccati equation is a consequence of the strong continuity of the transformation $h \to p_s$.

5. Control on the infinite set. In this section we shall develop a treatment of the control problem, Problem 2.1, on the infinite set $\sigma = \sigma_\infty = \{0, 1, 2, \cdots\}$.

Let $l^2(0, \infty; H)$ be the family of all sequences $\{x_i\}_{i\in\sigma}$ with $x_i \in H$ such that $\sum_\sigma \|x_i\|^2 < \infty$.

Remark 5.1 [6, p. 257]. $l^2(0, \infty; H)$ is a Hilbert space with the usual addition and scalar multiplication, and with an inner product defined by: for $x = \{x_i\}$ and $y = \{y_i\}$ in $l^2(0, \infty, H)$,

$$\langle x, y\rangle_{l^2(0,\infty;H)} = \sum_\sigma \langle x_i, y_i\rangle_H. \tag{5.1}$$

Similarly, $l^2(0, \infty; U)$ is a Hilbert space with inner product analogous to (5.1).

DEFINITION 5.1. A control u defined on σ is said to be *admissible* if $u \in l^2(0, \infty; U)$. A state x is said to be a *solution of* (2.1) *on* σ if it satisfies (2.1) with an admissible control u and $x_0 \in H$, and if $x(u) \in l^2(0, \infty; H)$, where $x(u) = \{x(u)_i\}_{i\in\sigma}$.

Notations. The state of the system will be denoted by x^N (to emphasize the dependence on N), that is, x^N is the solution of (2.1) on $\sigma_N = \{0, 1, 2, \cdots, N\}$. The cost functional in (2.2) is denoted by $J_N(u)$. Let (x^N, p^N) be the unique solution pair of (4.1) on σ_N and let K_i^N be the corresponding operator K_i in (4.11).

HYPOTHESIS 1. For every $u \in l^2(0, \infty; U)$ and $x_0 \in H$, there exists a unique solution of (2.1) on σ.

LEMMA 5.1. *Hypothesis* 1 *implies that the mapping* $u \to x$ *defined by the difference equation* (2.1) *is continuous from* $l^2(0, \infty; U)$ *into* $l^2(0, \infty; H)$.

Proof. Let T be the mapping $u \to x(u)$. Clearly T is linear. It remains to show that T is bounded. Define $T_n(u) = \{x_1(u), x_2(u), \cdots, x_n(u)\} \in l^2(0, \infty; H)$. Clearly T_n is linear. Assuming $x_0 = 0$ and by a simple calculation we can show that T_n is bounded for all n, i.e., $\|T_n(u)\| \leqq C_n\|u\|$ for all $u \in l^2(0, \infty; U)$ and for all n, where C_n is a constant which depends on n. Since for each $u \in l^2(0, \infty; U)$, $\|T_n(u)\|$ is bounded by a constant $\|x(u)\|$ for all n, by the uniform boundedness theorem there is a constant C such that $\|T_n\| \leqq C$ for all n. Since T is the strong limit of the sequence $\{T_n\}$, by Corollary 2 in [7, p. 69], T is a bounded linear operator.

Similar hypotheses have been used in the continuous-time problem; for example, see [9] and [10].

Remark 5.2. Hypothesis 1 implies that $x \in l^2(0, \infty; H)$ and $J_\infty(u) < \infty$ for all $u \in l^2(0, \infty; U)$.

THEOREM 5.1. *Assume that Hypothesis* 1 *holds. Then Problem* 2.1 *has a unique optimal control* $u^\infty \in l^2(0, \infty; U)$.

Proof. Because of Lemma 5.1, Lemma 3.1 can be used and the proof is identical to that of Theorem 3.1.

We now give a sufficient condition for Hypothesis 1 to hold.

THEOREM 5.2. *If* $\|\Phi\| < 1$, *then Hypothesis* 1 *will hold true.*

Proof. We are concerned with showing the existence and uniqueness of the solution $x(u) \in l^2(0, \infty; H)$ satisfying

$$x_{i+1} = \Phi x_i + Du_i; \qquad x_0 \in H. \tag{5.2}$$

Existence. Let $\|\Phi\| = a < 1$. Let $x(0)^N \in l^2(0, N; H)$ be the solution of (2.1) with zero control and $\hat{x}(u)^N \in l^2(0, N; H)$ be the solution of (2.1) corresponding to a control $u \in l^2(0, \infty; U)$ and $x_0 = 0$. Then the solution of (2.1) can be written in the form

$$x(u)^N = x(0)^N + \hat{x}(u)^N. \tag{5.3}$$

Now for $x(0)^N$, we derive

$$\begin{aligned}
0 &= \sum_{i=0}^{N-1} \langle x(0)_{i+1}^N, x(0)_{i+1}^N \rangle - \sum_{i=0}^{N-1} \langle \Phi x(0)_i^N, x(0)_{i+1}^N \rangle \\
&\geqq \sum_{i=0}^{N-1} \|x(0)_{i+1}^N\|^2 - \sum_{i=0}^{N-1} \|\Phi x(0)_i^N\| \, \|x(0)_{i+1}^N\| \\
&\geqq \sum_{i=0}^{N} \|x(0)_i^N\|^2 - \|x_0\|^2 - a\left(\sum_{i=0}^{N-1} \|x(0)_i^N\|^2 \sum_{i=0}^{N-1} \|x(0)_{i+1}^N\|^2 \right)^{1/2} \\
&\geqq \sum_{i=0}^{N} \|x(0)_i^N\|^2 - \|x_0\|^2 - a \sum_{i=0}^{N} \|x(0)_i^N\|^2 \\
&= (1 - a) \sum_{i=0}^{N} \|x(0)_i^N\|^2 - \|x_0\|^2,
\end{aligned}$$

hence,

$$\|x(0)^N\|_{l^2(0,N;H)} \leqq C.^1 \tag{5.4}$$

Similarly for $\hat{x}(u)^N$ with $x_0 = 0$ we derive

$$\sum_{i=0}^{N-1} \langle \hat{x}(u)_{i+1}^N, \hat{x}(u)_{i+1}^N \rangle - \sum_{i=0}^{N-1} \langle \Phi \hat{x}(u)_i^N, \hat{x}(u)_{i+1}^N \rangle = \sum_{i=0}^{N-1} \langle Du_i, \hat{x}(u)_{i+1}^N \rangle,$$

or

$$\sum_{i=0}^{N-1} \|\hat{x}(u)_{i+1}^N\|^2 - \sum_{i=0}^{N-1} \|\Phi \hat{x}(u)_i^N\| \, \|\hat{x}(u)_{i+1}^N\| \leqq \sum_{i=0}^{N-1} \|Du_i\| \, \|\hat{x}(u)_{i+1}^N\|,$$

or

$$\begin{aligned}
&\sum_{i=0}^{N-1} \|\hat{x}(u)_{i+1}^N\|^2 - a\left(\sum_{i=0}^{N-1} \|\hat{x}(u)_i^N\|^2 \sum_{i=0}^{N-1} \|\hat{x}(u)_{i+1}^N\|^2 \right)^{1/2} \\
&\qquad \leqq \left(\sum_{i=0}^{N-1} \|Du_i\|^2 \sum_{i=0}^{N-1} \|\hat{x}(u)_{i+1}^N\|^2 \right)^{1/2},
\end{aligned}$$

[1] The C's denoting constants independent of N.

or

$$\sum_{i=0}^{N-1} \|\hat{x}(u)_{i+1}^{N}\|^2 - a \sum_{i=0}^{N-1} \|\hat{x}(u)_{i+1}^{N}\|^2 \leqq \left(\sum_{i=0}^{N-1} \|Du_i\|^2\right)^{1/2} \left(\sum_{i=0}^{N-1} \|\hat{x}(u)_{i+1}^{N}\|^2\right)^{1/2},$$

or

$$(1-a)^2 \sum_{i=0}^{N-1} \|\hat{x}(u)_{i+1}^{N}\|^2 \leqq \sum_{i=0}^{N-1} \|Du_i\|^2 \leqq \sum_{i=0}^{\infty} \|Du_i\|^2.$$

Hence

$$\sum_{i=0}^{N-1} \|\hat{x}(u)_i^N\|^2 \leqq C_1 \sum_{i=0}^{\infty} \|u_i\|^2 \leqq C. \tag{5.5}$$

Therefore, from (5.3), (5.4) and (5.5), we have

$$\|x(u)^N\|_{l^2(0,N;H)} \leqq C. \tag{5.6}$$

Now if

$$\tilde{x}_i^N = \text{extension of } x(u)_i^N \text{ by 0 for } i > N,$$

$$\tilde{u}_i^N = \begin{cases} u_i & \text{in } [0, N), \\ 0 & \text{for } i \geqq n, \end{cases}$$

we have

$$\tilde{x}_{i+1}^N = \Phi \tilde{x}_i^N + D\tilde{u}_i^N - \Phi x_N^N \delta(i-N) \quad \text{on } [0, \infty), \tag{5.7}$$

where $\delta(0) = 1$, and $\delta(r) = 0$ for $r \neq 0$, and from (5.6),

$$\|\tilde{x}^N\|_{l^2(0,\infty;H)}^2 \leqq C. \tag{5.8}$$

We may then take a subsequence $N_n \to \infty$ such that

$$\tilde{x}^{N_n} \to x \quad \text{weakly in } l^2(0, \infty; H).$$

We then pass to the limit in (5.7), and hence

$$x_{i+1} = \Phi x_i + Du_i \quad \text{on } [0, \infty).$$

Uniqueness. Suppose x^1 and x^2 are solutions of (2.1) on σ with control $u \in l^2(0, \infty; U)$. Then

$$\|x_{i+1}^1 - x_{i+1}^2\| \leqq a\|x_i^1 - x_i^2\|. \tag{5.9}$$

Since $x_0^1 = x_0^2$, by iterating (5.9) we conclude that $x_i^1 = x_i^2$ for all $i \in \sigma$.

Thus we have shown that Hypothesis 1 holds.

Remark 5.3. The assumption that $\|\Phi\| < 1$ can be well-satisfied if $\Phi(t)$ is the semigroup of operators generated by a strongly elliptic operator. Because the operator $\Phi(t)$ has an exponential bound

$$\|\Phi(t)\| \leqq M e^{-\lambda t}, \qquad M, \lambda > 0,$$

and hence by a suitable choice of sampling interval $t = \delta$ (e.g., see Remark 2.3), we can always have $\|\Phi\| = \|\Phi(\delta)\| < 1$.

THEOREM 5.3. *Suppose* $\|\Phi\| < 1$; *then the adjoint state* p *is defined in a unique manner by*

$$p_i = \Phi^* p_{i+1} + Qx_i, \tag{5.10}$$

and

$$p \in l^2(0, \infty; H). \tag{5.11}$$

Proof. Let us simplify the notation by defining $f_i = Qx_i \in l^2(0, \infty; H)$. Then we are concerned with showing the existence and uniqueness of $p \in l^2(0, \infty; H)$ satisfying

$$p_i = \Phi^* p_{i+1} + f_i. \tag{5.12}$$

We note that $p \in l^2(0, \infty; H)$ implies that $p_\infty = 0$.

Uniqueness. Let $\|\Phi\| = a < 1$. Suppose (5.12) holds with $f = 0$. Then

$$\begin{aligned} 0 &= \sum_{i=0}^{\infty} \langle p_i, p_i \rangle - \sum_{i=0}^{\infty} \langle \Phi^* p_{i+1}, p_i \rangle \\ &\geqq \sum_{i=0}^{\infty} \|p_i\|^2 - \sum_{i=0}^{\infty} \|\Phi^* p_{i+1}\| \, \|p_i\| \\ &\geqq \sum_{i=0}^{\infty} \|p_i\|^2 - \left(\sum_{i=0}^{\infty} \|\Phi^* p_{i+1}\|^2 \sum_{i=0}^{\infty} \|p_i\|^2 \right)^{1/2} \\ &\geqq \sum_{i=0}^{\infty} \|p_i\|^2 - a \left(\sum_{i=0}^{\infty} \|p_{i+1}\|^2 \sum_{i=0}^{\infty} \|p_i\|^2 \right)^{1/2} \\ &\geqq \sum_{i=0}^{\infty} \|p_i\|^2 - a \sum_{i=0}^{\infty} \|p_i\|^2 \\ &= (1 - a) \sum_{i=0}^{\infty} \|p_i\|^2, \end{aligned} \tag{5.13}$$

hence $p = 0$.

Existence. Let q^N be the solution in $l^2(0, N; H)$ of

$$q_i^N = \Phi^* q_{i+1}^N + f_i \quad \text{in } [0, N), \tag{5.14}$$

where $q_N^N = 0$. Then

$$\sum_{i=0}^{N-1} \langle q_i^N, q_i^N \rangle - \sum_{i=0}^{N-1} \langle \Phi^* q_{i+1}^N, q_i^N \rangle = \sum_{i=0}^{N-1} \langle f_i, q_i^N \rangle. \tag{5.15}$$

With similar arguments for (5.13), we deduce from (5.15),

$$(1 - a) \sum_{i=0}^{N-1} \|q_i^N\|^2 \leqq \left(\sum_{i=0}^{N-1} \|f_i\|^2 \sum_{i=0}^{N-1} \|q_i^N\|^2 \right)^{1/2}$$

or

$$(1 - a)^2 \sum_{i=0}^{N-1} \|q_i^N\|^2 \leqq \sum_{i=0}^{N-1} \|f_i\|^2 \leqq \sum_{i=0}^{\infty} \|f_i\|^2.$$

Hence

$$\sum_{i=0}^{N-1} \|q_i^N\|^2 \leqq C_1 \sum_{i=0}^{\infty} \|f_i\|^2 \leqq C_2 \quad \text{(independent of } N\text{)}. \tag{5.16}$$

Now if

$$\tilde{q}_i^N = \text{extension of } q_i^N \text{ by 0 for } i > N,$$

$$\tilde{f}_i^N = \begin{cases} f_i & \text{in } [0, N), \\ 0 & \text{for } i \geqq N, \end{cases}$$

we have

$$\tilde{q}_i^N = \Phi^* \tilde{q}_{i+1}^N + \tilde{f}_i^N \quad \text{on } [0, \infty), \tag{5.17}$$

and hence from (5.16),

$$\sum_{i=0}^{\infty} \|\tilde{q}_i^N\|^2 \leqq C_2. \tag{5.18}$$

We may then take a subsequence $N_n \to \infty$ such that

$$\tilde{q}^{N_n} \to p \quad \text{weakly in } l^2(0, \infty; H).$$

We then pass to the limit in (5.17), and hence,

$$p_i = \Phi^* p_{i+1} + f_i \quad \text{on } [0, \infty).$$

Now we state the necessary condition for the optimal control on the infinite interval.

THEOREM 5.4. *Let* $\|\Phi\| < 1$. *If* $u^\infty \in l^2(0, \infty; U)$ *is the optimal control for Problem* 2.1 *with optimal response* $x^\infty \in l^2(0, \infty; H)$, *then there necessarily exists a unique adjoint state* $p^\infty \in l^2(0, \infty; H)$ *such that*

$$u_i^\infty = -R^{-1} D^* p_{i+1}^\infty, \tag{5.19}$$

$$p_i^\infty = \Phi^* p_{i+1}^\infty + Q x_i^\infty, \tag{5.20}$$

where the optimal response $x^\infty \in l^2(0, \infty; H)$ *satisfies*

$$x_{i+1}^\infty = \Phi x_i^\infty + D u_i^\infty; \qquad x_0^\infty = x_0. \tag{5.21}$$

Proof. The proof is identical to that of Theorem 3.2; hence it is omitted here.

THEOREM 5.5. *Let* $\|\Phi\| < 1$. *The optimal control* $u^\infty \in l^2(0, \infty; U)$ *for Problem* 2.1 *on the infinite interval is given by*

$$u_i^\infty = -R^{-1} D^* K^\infty (I + D R^{-1} D^* K)^{-1} \Phi x_i^\infty, \tag{5.22}$$

where K^∞ *is the bounded, positive semidefinite, self-adjoint operator, independent of* $i \in \sigma$, *satisfying the algebraic operator equation*

$$K^\infty = \Phi^* K^\infty (I + D R^{-1} D^* K^\infty)^{-1} \Phi + Q. \tag{5.23}$$

Moreover, the minimum cost using the optimal control (5.22) *on the interval* $[s, \infty)$ *is given by*

$$J_{s,\infty} = \langle K^\infty x_s^\infty, x_s^\infty \rangle. \tag{5.24}$$

Proof. If s is any fixed integer in $[0, \infty)$ and $x_s = h \in H$, then, using the same arguments as in the proofs of Lemma 4.1 and Corollary 4.1, it can be shown that the transformation $h \to p_s$ is continuous from H into H so that we can write

$$p_s^\infty = K_s^\infty h, \tag{5.25}$$

or, by using the fact that $s \in [0, \infty)$ is arbitrary, we have

$$p_i^\infty = K_i^\infty x_i^\infty \quad \text{for all } i \in [0, \infty), \tag{5.26}$$

where (x^∞, p^∞) is a unique solution pair of (5.20) and (5.21). Using the same arguments as in the proof of Theorem 4.1, we derive the Riccati-type equation

$$K_i^\infty = \Phi^* K_{i+1}^\infty (I + DR^{-1}D^* K_{i+1}^\infty)^{-1}\Phi + Q. \tag{5.27}$$

To prove that $K_i^\infty = K$ is independent of i. The system (5.20) and (5.21) is autonomous and is independent of the initial time s in the sense that if we translate the origin such that $i \to i - s$, then $p_s^\infty = p_0^\infty$. Hence from (5.25) we conclude that $K_s^\infty = K_0^\infty$.

The rest of the proof is identical to that of Theorem 4.2.

Let u^N and u^∞ be the optimal controls on $[0, N)$ and $[0, \infty)$, respectively. Now we shall examine the behavior as $N \to \infty$.

THEOREM 5.6. *Let $\|\Phi\| < 1$. Let $\tilde{u}^N$ ($\tilde{x}^N$, $\tilde{p}^N$ resp.) be the extension of u^N (x^N, p^N resp.) on $[0, \infty)$ by 0 outside $[0, N)$. Then as $N \to \infty$,*

$$\tilde{u}^N \to u^\infty \quad \textit{weakly in } l^2(0, \infty; U), \tag{5.28}$$

$$\tilde{x}^N \to x^\infty \quad \textit{weakly in } l^2(0, \infty; H), \tag{5.29}$$

$$\tilde{p}^N \to p^\infty \quad \textit{weakly in } l^2(0, \infty; H), \tag{5.30}$$

$$K_s^N h \to K^\infty h \quad \textit{weakly in } H, \quad \textit{for all } s \in \sigma, \quad \textit{for all } h \in H. \tag{5.31}$$

Proof. Let $j_N = \inf J_N(v)$, $j_\infty = \inf J_\infty(v)$. For $v \in l^2(0, \infty; U)$, we have $J_N(v) \leqq J_\infty(v)$ and hence $j_N \leqq j_\infty$. Thus $j_N = J_N(u^N) \geqq C \sum_{i=0}^{N-1} \|u_i^N\|^2$ and if $\tilde{u}^N$ is defined as in the statement of the theorem, by Remark 5.2 we have

$$\|\tilde{u}^N\|_{l^2(0,\infty;U)} \leqq C.^2 \tag{5.32}$$

But then due to Lemma 5.1 and (5.18),

$$\|\tilde{x}^N\|_{l^2(0,\infty;H)} \leqq C, \tag{5.33}$$

$$\|\tilde{p}^N\|_{l^2(0,\infty;H)} \leqq C, \tag{5.34}$$

and again by virtue of (5.33),

$$\|x_N^N\| \leqq C. \tag{5.35}$$

Hence we have from (4.1),

$$\tilde{p}_i^N = \Phi^* \tilde{p}_{i+1}^N + Q\tilde{x}_i^N - Qx_N^N \delta(i - N), \tag{5.36}$$

$$\tilde{x}_{i+1}^N = \Phi\tilde{x}_i^N - DR^{-1}D^*\tilde{p}_{i+1}^N - \Phi x_N^N \delta(i - N), \tag{5.37}$$

where $\delta(0) = 1$ and $\delta(r) = 0$ for $r \neq 0$.

[2] The C's denoting constants independent of N.

We may then find a sequence $N_n \to \infty$ such that

$$\tilde{u}^{N_n} \to \bar{u} \quad \text{weakly in } l^2(0, \infty; U),$$

$$\tilde{x}^{N_n} \to \bar{x}, \quad \tilde{p}^{N_n} \to \bar{p} \quad \text{weakly in } l^2(0, \infty; H).$$

Thus (5.36) and (5.37) become

$$\bar{p}_i = \Phi^* \bar{p}_{i+1} + Q\bar{x}_i, \tag{5.38}$$

$$\bar{x}_{i+1} = \Phi \bar{x}_i - DR^{-1}D^* \bar{p}_{i+1}. \tag{5.39}$$

By comparing the above with (5.19), (5.20) and (5.21), we deduce that $\bar{p} = p^\infty$ and $\bar{x} = x^\infty$. The relation

$$u_i^N = -R^{-1}D^* p_{i+1}^N \tag{5.40}$$

gives us, in the limit (if necessary, take a subsequence),

$$\bar{u}_i = -R^{-1}D\bar{p}_{i+1}; \tag{5.41}$$

hence we have, by comparing with (5.19), that $\bar{u} = u^\infty$. Thus we have proved (5.28), (5.29) and (5.30).

To prove (5.31), we observe that $K_s^N h$ is defined by

$$\begin{aligned} x_{i+1}^N &= \Phi x_i^N - DR^{-1}D^* p_{i+1}^N \quad \text{in } [s, N), \\ p_i^N &= \Phi p_{i+1}^N + Qx_i^N \quad \text{in } [s, N), \\ x_s^N &= h, \quad p_N^N = 0, \end{aligned} \tag{5.42}$$

and then

$$K_s^N h = p_s^N.$$

But (x^N, p^N) corresponds to the optimal control of a system whose state is given by

$$x_{i+1} = \Phi x_i + Du_i \quad \text{in } [s, N), \qquad x_s = h,$$

and whose cost is given by

$$J_{s,N}^h(u) = \sum_{i=s}^{N-1} [\langle x_i, Qx_i \rangle + \langle u_i, Ru_i \rangle].$$

By Theorem 4.2, we have

$$\inf J_{s,N}^h(u) = \langle K_s^N h, h \rangle \leqq J_{s,N}^h(0) \leqq C\|h\|^2;$$

hence

$$\|K_s^N h\| \leqq C\|h\|, \quad C = \text{const. independent of } s \text{ and } N. \tag{5.43}$$

Now if w^N is the optimal control of this problem, we obtain from (5.32),

$$\sum_{i=s}^{N-1} \|w_i^N\|^2 \leqq C,$$

and extending w^N by 0 for $i \geqq N$, we deduce that $\tilde{w}^N$ (resp. $\tilde{x}^N$, $\tilde{p}^N$) ranges in a bounded set of $l^2(0, \infty; U)$ (resp. $l^2(0, \infty; H)$). We may then find a sequence

$N_n \to \infty$ such that $\tilde{w}^{N_n} \to w$, $\tilde{x}^{N_n} \to \bar{x}$, $\tilde{p}^{N_n} \to \bar{p}$ in the corresponding weak topologies and hence satisfies (5.19), (5.20) and (5.21). Hence $\bar{x} = x^\infty$, $\bar{p} = p^\infty$ and $p_s^{N_n} \to p_s^\infty$ weakly in H, completing the proof.

Remark 5.3. Inclusion of the terminal time cost and the regulation from a desired state may be considered. Time-varying systems and cost functionals can be treated in the same framework.

REFERENCES

[1] A. G. BUTKOVSKIY, *Distributed Control Systems*, American Elsevier, New York, 1969.

[2] H. HALKIN, *Optimal control for systems described by difference equations*, Advances in Control Systems, vol. 1, Academic Press, New York, 1964, pp. 173–196.

[3] ———, *A maximum principle of the Pontryagin type for systems described by nonlinear difference equations*, this Journal, 4 (1966), pp. 90–111.

[4] D. KLEINMAN AND M. ATHANS, *The discrete minimum principle with application to the linear regulator problem*, Rep. ESL-R-260, M.I.T., Cambridge, Mass., 1966.

[5] J. L. LIONS, *Control optimal de systems gouvernés par des equations aux dériées partielles*, Dunod and Gauthier-Villars, Paris, 1968; English transl., Springer-Verlag, New York, 1971.

[6] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, John Wiley, New York, 1963.

[7] K. YOSIDA, *Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1968.

[8] K. Y. LEE, *Optimal sampled-data control of distributed parameter systems*, Doctoral thesis, Michigan State Univ., East Lansing, 1971.

[9] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[10] R. DATKO, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

A PROCEDURE FOR SOLVING CERTAIN DUAL OPTIMAL CONTROL PROBLEMS*

EARL R. BARNES†

Abstract. In this paper we describe an iterative procedure for solving a class of optimal control problems having quadratic cost functionals and linear isoperimetric equality constraints. An example is described in the Introduction which exhibits most of the constraints we are capable of handling. In a forthcoming paper [9] we show how the method applies to a fixed endpoint problem for a control system with lumped parameters, and give some numerical results. The present discussion is for systems with distributed parameters.

1. Introduction. Consider the following physical situation. We are given a container which is resting on a support that contacts only a small area of the base of the container. We wish to place a heaping load in the container in the most stable fashion. For stability it is clear that the center of gravity should be directly above the point of support, and as low as possible. But this is not enough. For it is easy to show, using the calculus of variations, that if the top surface of the load is not planar, the load can be rearranged slightly so as to lower its center of gravity. Moreover, if the load consists of particles without strong adhesive forces joining them, then the top surface of the load becomes unstable as the center of gravity approaches its greatest lower bound. Therefore, in trying to stabilize the load by lowering its center of gravity, we must take into account the stability of the top surface of the load. This surface will be in a stable equilibrium when its potential energy is a minimum. We take this into account in the mathematical formulation of our problem, to which we turn shortly. First we wish to inform the reader that the solution of this problem is not our primary concern here. However, we discuss the problem in considerable detail since its formulation exhibits most of the constraints that come up in our more general discussion.

Let Ω be a domain (open connected set) in the x, y-plane.[1] Consider a solid of

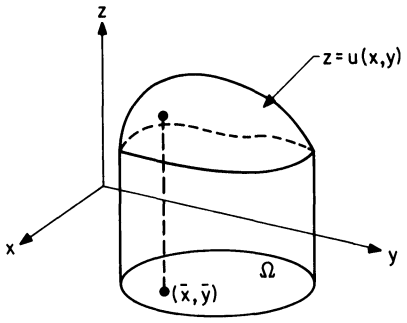


FIG. 1

* Received by the editors November 9, 1971, and in revised form November 29, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Mathematics Research Center, The University of Wisconsin, Madison, Wisconsin 53706. This research was supported by the U.S. Army under Contract DA-31-124-ARO-D-462.

[1] Actually, we assume that Ω is a normal domain. This means that in addition, Ω is bounded, and the Gauss integral theorem applies on Ω.

fixed volume v, bounded by Ω, a cylinder constructed on the boundary $\partial\Omega$ of Ω, and a surface $z = u(x, y)$. Let a point $(x_c, y_c) \in \Omega$ be given. Suppose it is required to determine the shape of the surface $z = u(x, y)$ so that the center of gravity of the solid will be along the line $x = x_c, y = y_c$ as close to the x, y-plane as possible, and such that the potential energy of the surface is a minimum. Of course, both of these objectives cannot be accomplished simultaneously. However, we can minimize the sum of the two approximately, and we shall do this. We assume the density of the solid is constant throughout. The coordinates of the center of gravity are then given by

$$x_c = \frac{\int\int_\Omega xu(x, y)\,dx\,dy}{\int\int_\Omega u(x, y)\,dx\,dy}, \qquad y_c = \frac{\int\int_\Omega yu(x, y)\,dx\,dy}{\int\int_\Omega u(x, y)\,dx\,dy},$$

$$z_c = \frac{\frac{1}{2}\int\int_\Omega u^2(x, y)\,dx\,dy}{\int\int_\Omega u(x, y)\,dx\,dy}.$$

In order to determine the potential energy of the surface let $\sqrt{1 + u_x^2 + u_y^2}\,dx\,dy$ denote an element of area on the surface. If the mass per unit area is denoted by σ, and the acceleration due to gravity by g, then the potential energy of the surface is given by

$$\sigma g \int\int_\Omega u(x, y)\sqrt{1 + u_x^2 + u_y^2}\,dx\,dy. \tag{1.1}$$

If we recall that the volume $\int\int_\Omega u(x, y)\,dx\,dy = v$ of our solid is fixed, we can give our problem the following variational formulation:

$$\text{minimize} \quad \int\int_\Omega \tfrac{1}{2}u^2(x, y) + u(x, y)\sqrt{1 + u_x^2 + u_y^2}\,dx\,dy \,^2 \tag{1.2}$$

subject to

$$\begin{aligned} &\int\int_\Omega u(x, y)\,dx\,dy = v, \qquad \int\int_\Omega xu(x, y)\,dx\,dy = vx_c, \\ &\int\int_\Omega yu(x, y)\,dx\,dy = vy_c \quad \text{and } u = \varphi \quad \text{on } \partial\Omega. \end{aligned} \tag{1.3}$$

The function φ describes the top edge of the container in Fig. 1. We assume that φ is continuous, and that φ can be extended to Ω in such a way that it is C^2 in Ω. We know that

$$\int\int_\Omega u\sqrt{1 + u_x^2 + u_y^2}\,dx\,dy \leqq \int\int_\Omega \tfrac{1}{2}u^2 + \tfrac{1}{2}(1 + u_x^2 + u_y^2)\,dx\,dy,$$

and because of this, we shall replace the cost functional (1.2) by the more manageable functional

$$\int\int_\Omega u^2 + u_x^2 + u_y^2\,dx\,dy. \tag{1.4}$$

[2] At this point we feel free to attach weights to the potential energy, and the z-coordinate of the center of gravity so as to keep the notation simple. This will all be justified in § 2.

Without this, or a similar, simplification, we would be led to difficult questions in nonlinear partial differential equations. Almost no control theory has been developed for such systems. We shall make other simplifying assumptions, but as we shall see in § 5, they can be dispensed with.

According to Lagrange's multiplier rule, we can obtain the minimum of (1.4), subject to the constraints (1.3), by minimizing the functional

$$\int\int_{\Omega} u^2 + u_x^2 + u_y^2 - \eta_1 u - \eta_2 x u - \eta_3 y u \, dx \, dy \tag{1.5}$$

(for some appropriate choice of the multipliers η_1, η_2, η_3) over the class of functions which are C^2 in Ω and equal to φ on $\partial\Omega$.

The function which minimizes (1.5) subject to (1.3) is known to satisfy the boundary value problem (Euler–Lagrange equation)

$$\begin{aligned} &u_{xx} + u_{yy} - u = \tfrac{1}{2}(\eta_1 + \eta_2 x + \eta_3 y), && (x, y) \in \Omega, \\ &u = \varphi \quad \text{on } \partial\Omega. \end{aligned} \tag{1.6}$$

This boundary value problem has a solution for any choice of the multipliers η_1, η_2, η_3 (cf. [1, Theorem 2, p. 177]). Actually, this theorem refers to the case where $\varphi = 0$ on $\partial\Omega$, but it can be made to apply to the present situation as follows. Assume φ has been extended to be C^2 in Ω. Write $u = w + \varphi$. Then by the theorem just referenced, we can determine w satisfying

$$w_{xx} + w_{yy} - w = \tfrac{1}{2}(\eta_1 + \eta_2 x + \eta_3 y) - (\varphi_{xx} + \varphi_{yy} - \varphi)$$

on Ω, $w = 0$ on $\partial\Omega$. With w so determined, u satisfies (1.6).

To obtain the solution of our variational problem, we solve (1.6) with the η_i appearing as parameters. Proper values of these parameters are then determined by the conditions (1.3).

In applied mathematics it is quite common to approximate the energy in the surface of our system by one half the quantity (1.4). Then, in the special case where $u = 0$ on $\partial\Omega$, we have, by the Gauss integral theorem,

$$\begin{aligned} \int\int_{\Omega} u^2 + u_x^2 + u_y^2 \, dx \, dy &= \int\int_{\Omega} u(u - u_{xx} - u_{yy}) \, dx \, dy \\ &\leqq \left(\int\int_{\Omega} u^2 + u_x^2 + u_y^2 \, dx \, dy \right)^{1/2} \\ &\quad \left(\int\int_{\Omega} (u - u_{xx} - u_{yy})^2 \, dx \, dy \right)^{1/2} \end{aligned} \tag{1.7}$$

from which we deduce the so-called energy inequality

$$\int\int_{\Omega} u^2 + u_x^2 + u_y^2 \, dx \, dy \leqq \int\int_{\Omega} (u - u_{xx} - u_{yy})^2 \, dx \, dy. \tag{1.8}$$

In problems of optimal control, it is often desired to minimize the expression on the right in (1.8), and these problems are referred to as "minimum energy" control problems.

2. An optimal control problem. Let us now approximate the energy in our surface by

$$\frac{1}{2}\int\int_{\Omega}(u - u_{xx} - u_{yy})^2\,dx\,dy.^3$$

For simplicity, we shall write

$$\begin{aligned} &u_{xx} + u_{yy} - u = f(x, y), && (x, y) \in \Omega, \\ &u = \varphi \quad \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

Recalling the expression for the vertical height of the center of gravity of our system, we see that our variational problem is now to find a function f such that the functional

$$\frac{1}{2}\int\int_{\Omega} u^2 + f^2\,dx\,dy$$

is minimized subject to (2.1), (1.3) and $u = 0$ on $\partial\Omega$. This problem can still be handled by methods in the calculus of variations. But at this point we place a constraint on f which takes our problem out of the realm of the calculus of variations and into the more modern theory of optimal control. This constraint will guarantee that our surface satisfies $u \geqq 0$, a condition without which our problem makes no sense. Let us assume that our surface has $u(x, y) < 0$ for some $(x, y) \in \Omega$. This cannot be a boundary point of Ω since $u = 0$ on $\partial\Omega$. It follows that u achieves its minimum value on Ω at a point where $u < 0$. But at a minimum point of u we have $u_{xx} \geqq 0$ and $u_{yy} \geqq 0$. It follows that

$$u_{xx} + u_{yy} - u > 0$$

at the point where u assumes its minimum value. It is now clear that we can satisfy the constraint $u \geqq 0$ by placing the constraint $f \leqq 0$ on f in (2.1). If we do this we obtain a problem which cannot be handled by the calculus of variations. The Euler–Lagrange equations must be replaced by the more powerful "maximum principle" and other adjustments must be made. For example, the Lagrange multipliers which appear as parameters in the maximum principle can no longer be determined with ease. However, we can determine them, and the solution to our problem, to any desired degree of accuracy by an iterative procedure to be described below. But before we come to this, we must develop the necessary theoretical background. We shall not strive for the greatest generality, since the notation is already complicated enough. However, we can be slightly more general than we have been until now.

We assume that our system is governed by the equation (2.1), where now, f is required simply to be a bounded measurable function on Ω, which assumes values in a given closed interval I. I may be finite or infinite. Such an f will be referred to as an admissible control, and we shall denote the class of admissible controls by $\mathscr{C}$. The problem of optimal control is to find an admissible control f, such that the pair (f, u) satisfying (2.1) minimizes the functional

$$c_0(f) = \int\int_{\Omega} p(x, y)u^2 + q(x, y)(u_x^2 + u_y^2) + r(x, y)f^2\,dx\,dy \tag{2.2}$$

[3] In doing this we tacitly assume that $\varphi = 0$.

subject to a set of isoperimetric constraints of the form

$$\int\int_{\Omega} \{a_{i0}(x, y)u + a_{i1}(x, y)u_x + a_{i2}(x, y)u_y + b_i(x, y)f\}\, dx\, dy = c_i, \tag{2.3}$$

$i = 1, 2, \cdots, n$. An $f \in \mathscr{C}$ which solves this problem will be called an optimal control. All the coefficients in (2.2) and (2.3) are assumed to be continuous, and the coefficients of u_x and u_y even C^1. Moreover, the coefficients in (2.2) are required to be nonnegative,[4] with the further stipulation that if a coefficient in (2.2) is identically zero, then the corresponding coefficients of that variable in (2.3) must also be identically zero. Thus if one of the variables u, u_x, u_y, f does not appear in (2.2) it must also be absent from (2.3). The c_i in (2.3) are arbitrary real constants, assumed given.

Before discussing the optimal control problem we must first say what we mean by a solution of (2.1) corresponding to an f which is only bounded and measurable. To this end, let us assume for the moment that f is continuous. Then it follows (as in the previous section) from [1, Theorem 2, p. 177] that the boundary value problem (2.1) has a solution which is C^2 in Ω and C^0 on $\partial\Omega$. Now for a general $f \in \mathscr{C}$, let $\{f_n\}$ be a sequence of continuous functions converging in the mean square to f. Corresponding to each f_n there is a classical solution u_n of our boundary value problem (2.1) with f replaced by f_n. Moreover, since any difference $u_n - u_k$ vanishes on $\partial\Omega$, we can apply the energy inequality (1.8) to these differences. This will show that the sequences $\{u_n\}$, $\{u_{nx}\}$, $\{u_{ny}\}$ are Cauchy with respect to the mean square norm. It follows that they converge in the mean square to square integrable functions, which we denote by u, u_x and u_y, respectively. We call u the weak solution of (2.1) corresponding to f. u_x, u_y are called generalized derivatives of u.

The Gauss integral theorem applies to weak solutions with zero boundary conditions since such weak solutions are limits, in the mean square, of classical solutions which vanish outside sets of compact support in Ω, and to which the Gauss integral theorem can be applied, the integral over the boundary never appearing. The result for weak solutions is obtained by passing to the limit in the result for classical solutions. We shall use this fact many times in the sequel, without justifying it each time.

THEOREM 2.1. *If there exists a control $f \in \mathscr{C}$ for which the constraints* (2.3) *are satisfied, then there exists an optimal control.*

Proof. This theorem follows from [10, Chap. 1, Theorem 1.1]. For a given $f \in \mathscr{C}$ we define a vector $a = (a_1, \cdots, a_n) \in E_n$, by

$$a_i = \int\int_{\Omega} a_{i0}u + a_{i1}u_x + a_{i2}u_y + b_i f\, dx\, dy, \tag{2.4}$$

$i = 1, \cdots, n$. And when a is used, the corresponding f will be understood from context. Similarly, an f, or a u, with a given subscript, or superscript, will be understood to be associated with the u, or f, having the same subscript, or superscript, by (2.1). Also, Δu will correspond to Δf.

[4] Here nonnegative means strictly positive on Ω, or identically zero there.

We denote by E the space of four-dimensional vector functions (u, v, w, z), where u, v, w, z are square integrable functions on Ω. An inner product is defined on E by

$$\langle Z_1, Z_2 \rangle = \int\int_{\Omega} pu_1u_2 + qv_1v_2 + qw_1w_2 + rz_1z_2 \, dx \, dy,$$

where $Z_i = (u_i, v_i, w_i, z_i)$, $i = 1, 2$. The norm of a vector $Z \in E$ is defined by

$$\|Z\| = \sqrt{\langle Z, Z \rangle}.$$

With these conventions adopted, consider the set of attainability

$$A = \{(c_0(f), a) | f \in \mathscr{C}\}.$$

The problem of optimal control is to find the point $(c_0(f), a)$ in A for which $a = c$ and $c_0(f)$ is as small as possible. Let $\tilde{A}$ denote the saturation set of A, defined as follows (cf. [2]).

DEFINITION 2.1. $\tilde{A}$ is the set of points $(\tilde{c}_0, \tilde{a}) \in E_{n+1}$ for which there exists a point $(c_0(f), a) \in A$ with $a = \tilde{a}$ and $c_0(f) \leqq \tilde{c}_0$.

It is an easy matter to show that $\tilde{A}$ is convex (cf. [2]). Moreover, if f^* is an optimal control, $(c_0(f^*), c)$ is a boundary point of $\tilde{A}$. This fact allows a geometric interpretation of the next theorem (maximum principle).

THEOREM 2.2. *In order that the admissible control f^* be optimal, it is necessary that there exist constants $\eta_0 \leqq 0, \eta_1, \cdots, \eta_n$, not all zero, and a function v satisfying*

$$\begin{aligned} v_{xx} + v_{yy} - v = {} & \sum_{i=1}^{n} \eta_i(a_{i0} - \partial a_{i1}/\partial x - \partial a_{i2}/\partial y) \\ & + 2\eta_0\{(p - q)u^* - q_x u_x^* - q_y u_y^* - qf^*\} \quad \text{on } \Omega, \\ & v = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.5}$$

such that for almost all $(x, y) \in \Omega$,

$$\begin{aligned} & \max_{f \in I} \{v(x, y)f + \sum_{i=1}^{n} \eta_i b_i(x, y)f + \eta_0 r(x, y)f^2\} \\ & \quad = v(x, y)f^*(x, y) + \sum_{i=1}^{n} \eta_i b_i(x, y)f^*(x, y) + \eta_0 r(x, y)f^{*2}(x, y). \end{aligned} \tag{2.6}$$

This theorem is essentially proved in [10, Chap. 2] and so we shall not prove it here. However, we do wish to give a geometric interpretation of the theorem.

Recall that the saturation set $\tilde{A}$ discussed above is convex and that $(c_0(f^*), c)$ is the lowest point in $\tilde{A}$ on the ray $a = c$. We picture $\tilde{A}$ here with the c_0 axis drawn vertically. Since $(c_0(f^*), c)$ is in the boundary of $\tilde{A}$, there exists a nonzero vector $(\eta_0, \eta) \in E_{n+1}$ which is normal to $\tilde{A}$ at $(c_0(f^*), c)$. For any point $(c_0(f), a)$ in $\tilde{A}$ we have

$$(\eta_0, \eta) \cdot ((c_0(f), a) - (c_0(f^*), c)) \leqq 0. \tag{2.7}$$

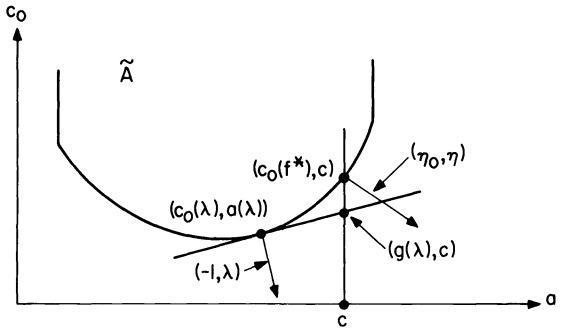


FIG. 2

This says that

$$\max_{f \in C} (\eta_0, \eta) \cdot (c_0(f), a) = (\eta_0, \eta) \cdot (c_0(f^*), c)$$

which is equivalent to (2.6).

3. Computing the optimal control: The dual problem. In this section we describe an iterative procedure for computing the optimal control f^*. We consider only the case which corresponds to so-called "normal" problems in the calculus of variations. These are problems where the multiplier η_0 in (2.7) must of necessity be <0. In this case we can without loss of generality assume that $\eta_0 = -1$, and we shall do this.

Let $\lambda \in E_n$ be arbitrary. Choose $f \in \mathscr{C}$ to minimize the functional

$$\begin{aligned} -(-1, \lambda) \cdot (c_0(f), a) = \int\int_\Omega & pu^2 + q(u_x^2 + u_y^2) + rf^2 \\ & - \sum_{i=1}^{n} \lambda_i (a_{i0}u + a_{i1}u_x + a_{i2}u_y + b_i f)\, dx\, dy \end{aligned} \tag{3.1}$$

subject only to the constraint (2.1). We denote the solution to this problem by $f(x, y, \lambda)$. The corresponding values of u will be denoted by $u(x, y, \lambda)$. The cost $c_0(f)$ of f will be denoted by $c_0(\lambda)$, and the corresponding point in $\tilde{A}$ will be denoted by $(c_0(\lambda), a(\lambda))$. Clearly $(c_0(\lambda), a(\lambda))$ is in the boundary of $\tilde{A}$ and the vector $(-1, \lambda)$ is normal to $\tilde{A}$ at $(c_0(\lambda), a(\lambda))$. (See Fig. 2.) Let $(g(\lambda), c)$ denote the point where the plane tangent to $\tilde{A}$ at $(c_0(\lambda), a(\lambda))$ intersects the line $a = c$. We have $(-1, \lambda) \cdot ((g(\lambda), c) - (c_0(\lambda), a(\lambda))) = 0$, so that

$$g(\lambda) = \lambda \cdot (c - a(\lambda)) + c_0(\lambda).$$

THEOREM 3.1. *The multiplier η appearing in the maximum principle has the property that $f(x, y, \eta) = f^*(x, y)$ and*

$$\max_{\lambda \in E_n} g(\lambda) = g(\eta). \tag{3.2}$$

Moreover, g is concave and C^1 on E_n, and the gradient of g is given by $\nabla g(\lambda) = c - a(\lambda)$.

We shall not prove this theorem here. Its proof can be deduced from the very general discussion in [4, Theorem 2.1] applied to the function

$$G(\lambda) = (1, -\lambda) \cdot (c_0(\lambda), a(\lambda)) = \min_{(c_0, \tilde{a}) \in \tilde{A}} (1, -\lambda) \cdot (\tilde{c}_0, \tilde{a}).$$

It can also be deduced from the very lucid discussion of duality for a minimum fuel problem in a lumped parameter model by Neustadt and Meditch in [5]. See also [6].

We shall make use of Theorem 3.1 to compute the optimal control f^*. We proceed as follows: A method is described for computing $f(x, y, \lambda)$ for any given $\lambda \in E_n$. Given $f(x, y, \lambda)$, we can then evaluate $g(\lambda)$ and $\nabla g(\lambda)$. We then give a method for constructing a sequence $\{\lambda_k\}$ satisfying

$$\lim_{k \to \infty} g(\lambda_k) = \max_{\lambda \in E_n} g(\lambda)$$

and

$$\lim_{k \to \infty} \int\int_{\Omega} (f(x, y, \lambda_k) - f^*(x, y))^2 \, dx \, dy = 0.$$

Thus we shall describe an iterative procedure for solving the problem (3.2). And in so doing, we obtain the solution to our control problem. This shows that problem (3.2) is equivalent to the optimal control problem. The two problems are said to be dual to each other.

We postpone, for the time being, the problem of how to compute $f(x, y, \lambda)$ for a given λ, and go immediately into the discussion of maximizing g, i.e., to the solution of the dual problem.

(a) Let β be a constant > 1, and let θ be a constant satisfying $0 < \theta < 4$.

(b) Let $\lambda_1 \in E_n$ be chosen arbitrarily.

(c) If $\lambda_1, \cdots, \lambda_k \in E_n$ have been chosen, take $\lambda_{k+1} = \lambda_k + \theta t_k s_k$, where $s_k = (s_k^1, \cdots, s_k^n)$ is any vector in E_n satisfying

$$|\nabla g(\lambda_k)|^2 \leqq s_k \cdot \nabla g(\lambda_k), \qquad |s_k|^2 \leqq \beta |\nabla g(\lambda_k)|^2$$

and

$$t_k = \frac{|\nabla g(\lambda_k)|^2}{\displaystyle\int\int_{\Omega} p^{-1} \left(\sum_{i=1}^{n} a_{i0} s_k^i \right)^2 + q^{-1} \left(\sum_{i=1}^{n} a_{i1} s_k^i \right)^2 + q^{-1} \left(\sum_{i=1}^{n} a_{i2} s_k^i \right)^2 + r^{-1} \left(\sum_{i=1}^{n} b_i s_k^i \right)^2 \cdot dx \, dy}.$$

THEOREM 3.2. *The sequence of admissible controls* $\{f(x, y, \lambda_k)\}$, *corresponding to the sequence* $\{\lambda_k\}$ *generated in* (a)–(c), *converges in the mean square to the optimal control* f^*.

For the proof of this theorem we need a lemma.

LEMMA 3.1. *Let λ and γ be any two vectors in E_n. Let $\Delta f = f(x, y, \gamma) - f(x, y, \lambda)$, $\Delta u = u(x, y, \gamma) - u(x, y, \lambda)$. Then*

$$\int\int_\Omega p(\Delta u)^2 + q(\Delta u_x)^2 + q(\Delta u_y)^2 + r(\Delta f)^2 \, dx\, dy$$

$$\leqq \frac{1}{4}\int\int_\Omega p^{-1}\left(\sum_{i=1}^n (\gamma_i - \lambda_i)a_{i0}\right)^2 + q^{-1}\left(\sum_{i=1}^n (\gamma_i - \lambda_i)a_{i1}\right)^2 \tag{3.3}$$

$$+ q^{-1}\left(\sum_{i=1}^n (\gamma_i - \lambda_i)a_{i2}\right)^2 + r^{-1}\left(\sum_{i=1}^n (\gamma_i - \lambda_i)b_i\right)^2 dx\, dy.$$

Proof. Let E denote the space introduced in § 2. The set

$$K = \{(u, u_x, u_y, f) | f \in \mathscr{C}\}$$

is clearly a convex subset of E.

Let us now observe that

$$\int\int_\Omega pu^2 + qu_x^2 + qu_y^2 + rf^2 - \sum_{i=1}^n \lambda_i(a_{i0}u + a_{i1}u_x + a_{i2}u_y + b_i f)\, dx\, dy$$

$$= \int\int_\Omega \left(p^{1/2}u - \tfrac{1}{2}p^{-1/2}\sum_{i=1}^n \lambda_i a_{i0}\right)^2 + \left(q^{1/2}u_x - \tfrac{1}{2}q^{-1/2}\sum_{i=1}^n \lambda_i a_{i1}\right)^2$$

$$+ \left(q^{1/2}u_y - \tfrac{1}{2}q^{-1/2}\sum_{i=1}^n \lambda_i a_{i2}\right)^2 + \left(r^{1/2}f - \tfrac{1}{2}r^{-1/2}\sum_{i=1}^n \lambda_i b_i\right)^2 dx\, dy$$

$$+ \text{const}.$$

Since $f(x, y, \lambda)$ minimizes this expression, it is clear that $(u(x, y, \lambda), u_x(x, y, \lambda), u_y(x, y, \lambda), f(x, y, \lambda))$ is the point in K nearest the point

$$\left(\tfrac{1}{2}p^{-1}\sum_{i=1}^n \lambda_i a_{i0}, \tfrac{1}{2}q^{-1}\sum_{i=1}^n \lambda_i a_{i1}, \tfrac{1}{2}q^{-1}\sum_{i=1}^n \lambda_i a_{i2}, \tfrac{1}{2}r^{-1}\sum_{i=1}^n \lambda_i b_i\right).$$

We know that the point in a convex set nearest a given point is the projection of the given point onto the convex set. Moreover, the projections of two given points onto a convex set are closer to each other than are the two points (cf. [7, Problem 1, p. 66]). Inequality (3.3) is simply a statement of this fact for the points corresponding to γ and λ above.

Proof of Theorem 3.2. For $k \geqq 1$ we have

$$g(\lambda_{k+1}) - g(\lambda_k) = \int_0^1 \frac{d}{d\tau} g(\lambda_k + \tau\theta t_k s_k)\, d\tau$$

$$= \int_0^1 [\nabla g(\lambda_k + \tau\theta t_k s_k) - \nabla g(\lambda_k)] \cdot \theta t_k s_k \, d\tau + \theta t_k s_k \cdot \nabla g(\lambda_k) \tag{3.4}$$

$$= -\int_0^1 [a(\lambda_k + \tau\theta t_k s_k) - a(\lambda_k)] \cdot \theta t_k s_k \, d\tau + \theta t_k s_k \cdot \nabla g(\lambda_k).$$

Now, putting $\gamma = \lambda_k + \tau\theta t_k s_k$, $\lambda = \lambda_k$, and using the obvious meanings of Δu and Δf, we have

$$
\begin{aligned}
[a(\gamma) - a(\lambda)] \cdot s_k &= \int\int_\Omega \sum_{i=1}^{n} (a_{i0}p^{-1/2}p^{1/2}\Delta u s_k^i + a_{i1}q^{-1/2}q^{1/2}\Delta u_x s_k^i \\
&\qquad + a_{i2}q^{-1/2}q^{1/2}\Delta u_y s_k^i + b_i r^{-1/2}r^{1/2}\Delta f_k^i s_k^i)\, dx\, dy \\
&\leqq \left(\int\int_\Omega p^{-1}\left(\sum_{i=1}^{n} a_{i0}s_k^i \right)^2 + q^{-1}\left(\sum_{i=1}^{n} a_{i1}s_k^i \right)^2 + q^{-1}\left(\sum_{i=1}^{n} a_{i2}s_k^i \right)^2 \right. \\
&\qquad \left. + r^{-1}\left(\sum_{i=1}^{n} b_i s_k^i \right)^2 dx\, dy \right)^{1/2} \\
&\qquad \cdot \left(\int\int_\Omega p(\Delta u)^2 + q(\Delta u_x)^2 + q(\Delta u_y)^2 + r(\Delta f)^2\, dx\, dy \right)^{1/2} \\
&\leqq \frac{\tau\theta t_k}{2} \int\int_\Omega p^{-1}\left(\sum_{i=1}^{n} a_{i0}s_k^i \right)^2 + q^{-1}\left(\sum_{i=1}^{n} a_{i1}s_k^i \right)^2 \\
&\qquad + q^{-1}\left(\sum_{i=1}^{n} a_{12}s_k^i \right)^2 + r^{-1}\left(\sum_{i=1}^{n} b_i s_k^i \right)^2 dx\, dy \\
&\qquad\qquad \text{(by Lemma 3.1)} \\
&= \frac{\tau\theta}{2} |\nabla g(\lambda_k)|^2.
\end{aligned}
$$

Going back to (3.4) we have

$$
\begin{aligned}
g(\lambda_{k+1}) - g(\lambda_k) &\geqq -\int_0^1 \frac{\tau\theta^2 t_k}{2} |\nabla g(\lambda_k)|^2\, d\tau + \theta t_k |\nabla g(\lambda_k)|^2 \\
&= \theta t_k (1 - \theta/4) |\nabla g(\lambda_k)|^2.
\end{aligned} \tag{3.5}
$$

By summing the right-hand side of this inequality from 1 to ∞ we have

$$
g(\eta) - g(\lambda_1) \geqq \sum_{k=1}^{\infty} \theta t_k (1 - \theta/4) |\nabla g(\lambda_k)|^2, \tag{3.6}
$$

and since clearly,

$$
t_k \geqq \left[\beta \int\int_\Omega p^{-1} \sum_{i=1}^{n} a_{i0}^2 + q^{-1} \sum_{i=1}^{n} (a_{i1}^2 + a_{i2}^2) + r^{-1} \sum_{i=1}^{n} b_i^2\, dx\, dy \right]^{-1},
$$

it follows from (3.6) that

$$
\lim_{k \to \infty} \nabla g(\lambda_k) = 0. \tag{3.7}
$$

Now consider the inner product $(\eta - \lambda_k) \cdot \nabla g(\lambda_k)$. This value tends to zero as $k \to \infty$. To show this, it suffices to show that the sequence $\{\lambda_k\}$ is bounded, by (3.7). Assume that some subsequence of $\{\lambda_k\}$, which we denote also by $\{\lambda_k\}$, is unbounded. Then since $f(x, y, \lambda_k)$ minimizes the functional (3.1) with $\lambda = \lambda_k$, we have

$$
\left(-\frac{1}{|\lambda_k|}, \frac{\lambda_k}{|\lambda_k|} \right) \cdot ((c_0(f), a) - (c_0(\lambda_k), a(\lambda_k))) \leqq 0 \tag{3.8}
$$

for all $f \in \mathscr{C}$. Let us assume, by selecting a subsequence if necessary, that the sequences $\{c_0(\lambda_k)\}$ and $\{\lambda_k/|\lambda_k|\}$ converge to points $\hat{c}_0 \in E_1$ and $\hat{\lambda} \in E_n$, respectively, and $\hat{\lambda} \neq 0$. Then since $\lim_{k\to\infty} a(\lambda_k) = c$, by (3.7), we can pass to the limit in (3.8) to obtain

$$(0, \hat{\lambda}) \cdot ((c_0(f), a) - (\hat{c}_0, c)) \leqq 0$$

for any $f \in \mathscr{C}$. But it is clear from this inequality that we can write

$$(0, \hat{\lambda}) \cdot ((c_0(f), a) - (c_0(f^*), c)) \leqq 0, \qquad f \in \mathscr{C}.$$

And this inequality contradicts our normality assumption. Therefore the sequence $\{\lambda_k\}$ is bounded.

Let us now write

$$\begin{aligned}(\eta - \lambda_k) \cdot \nabla g(\lambda_k) &= (\eta - \lambda_k) \cdot (\alpha(\eta) - a(\lambda_k)) \\ &= (-1, \lambda_k) \cdot ((c_0(\lambda_k), a(\lambda_k)) - (c_0(\eta), a(\eta))) \\ &\quad + (-1, \eta) \cdot ((c_0(\eta), a(\eta)) - (c_0(\lambda_k), a(\lambda_k))).\end{aligned} \tag{3.9}$$

Let v_1 and v_2 be two functions satisfying

$$\begin{aligned}v_{1xx} + v_{1yy} - v_1 = \sum_{i=1}^{n} \lambda_k^i \left(a_{i0} - \frac{\partial a_{i1}}{\partial x} - \frac{\partial a_{i2}}{\partial y} \right) \\ -2\{(p - q)u(x, y, \lambda) - q_x u_x(x, y, \lambda) - q_y u_y(x, y, \lambda) \\ - qf(x, y, \lambda)\}\end{aligned}$$

(λ_k^i is the ith component of λ_k) and

$$\begin{aligned}v_{2xx} + v_{2yy} - v_2 = \sum_{i=1}^{n} \eta_i \left(a_{i0} - \frac{\partial a_{i1}}{\partial x} - \frac{\partial a_{i2}}{\partial y} \right) \\ -2\{(p - q)u^* - q_x u_x^* - q_y u_y^* - qf^*\}\end{aligned}$$

on Ω, and $v_i = 0$ on $\partial\Omega$, $i = 1, 2$. Then by expanding the terms on the right in (3.9) and integrating by parts we obtain

$$\begin{aligned}(\eta - \lambda_k) \cdot \nabla g(\lambda_k) = \int\int_{\Omega} \Delta v \Delta f + \sum_{i=1}^{n} (\eta_i - \lambda_k^i) b_i \Delta f \, dx \, dy \\ + 2 \int\int_{\Omega} p(\Delta u)^2 + q(\Delta u_x)^2 + q(\Delta u_y)^2 \, dx \, dy,\end{aligned} \tag{3.10}$$

where $\Delta v = v_2 - v_1$, $\Delta f = f(x, y, \eta) - f(x, y, \lambda_k)$, $\Delta u = u(x, y, \eta) - u(x, y, \lambda_k)$. Note that $c = a(\eta)$ and $f^*(x, y) = f(x, y, \eta)$, etc. For convenience we shall write $f_1 = f(x, y, \lambda_k)$, $f_2 = f(x, y, \eta)$ in the next paragraph.

We have

$$\begin{aligned}\int\int_{\Omega}(v_2 - v_1)(f_2 - f_1) + \sum_{i=1}^{n}(\eta_i - \lambda_i)b_i(f_2 - f_1)\,dx\,dy \\ = \int\int_{\Omega} v_2 f_2 + \sum_{i=1}^{n}\eta_i b_i f_2 - r f_2^2 \\ - \left\{v_2 f_1 + \sum_{i=1}^{n}\eta_i b_i f_1 - r f_1^2\right\}dx\,dy \\ + \int\int_{\Omega} v_1 f_1 + \sum_{i=1}^{n}\lambda_k^i b_i f_1 - r f_1^2 \\ - \left\{v_1 f_2 + \sum_{i=1}^{n}\lambda_k^i b_i f_2 - r f_2^2\right\}dx\,dy.\end{aligned} \tag{3.11}$$

Fix (x, y) in Ω and consider the quadratic function

$$Q(f) = v_2(x, y)f + \sum_{i=1}^{n}\eta_i b_i(x, y)f - r(x, y)f^2.$$

According to the maximum principle (Theorem 2.2),

$$\max_{f\in I} Q(f) = Q(f^*(x, y)) = Q(f_2) \tag{3.12}$$

for almost every choice of $(x, y)\in\Omega$. If (x, y) is a point for which this condition holds, we have

$$Q'(f_2)(f - f_2) \leqq 0$$

for all $f\in I$. To see this, observe that the converse implies that for some $f\in I$,

$$Q(f_2 + \alpha(f - f_2)) - Q(f_2) = \alpha Q'(f_2)(f - f_2) + O(\alpha^2) > 0$$

for α sufficiently small. This contradicts (3.12). Considering the first integrand on the right in (3.11) we have

$$\begin{aligned}Q(f_1) - Q(f_2) &= Q'(f_2)(f_1 - f_2) - r(f_1 - f_2)^2 \\ &\leqq -r(f_1 - f_2)^2,\end{aligned}$$

so that $Q(f_2) - Q(f_1) \geqq r(f_1 - f_2)^2$. From this it follows that the first integral on the right in (3.11) is $\geqq \int\int_{\Omega} r(f_1 - f_2)^2\,dx\,dy$. If now we observe that the control $f_1 = f(x, y, \lambda_k)$ also solves an optimal control problem (it minimizes the functional (3.1)) similar to that solved by $f(x, y, \eta)$, we see that the exact same argument can be used to show that the second integral on the right in (3.11) is $\geqq \int\int_{\Omega} r(f_1 - f_2)^2\,dx\,dy$. If we now return to (3.10) we can write

$$(\eta - \lambda_k)\cdot\nabla g(\lambda_k) \geqq 2\int\int_{\Omega} p(\Delta u)^2 + q(\Delta u_x)^2 + q(\Delta u_y)^2 + r(\Delta f)^2\,dx\,dy.$$

It follows from this inequality, together with (3.7) and the boundedness of the sequence $\{\lambda_k\}$, that the sequence $\{f(x, y, \lambda_k)\}$ converges in the mean square to

$f^*(x, y)$. We have also shown that $u(x, y, \lambda_k)$ and its derivatives converge, and to the proper values, as $k \to \infty$. This completes the proof of Theorem 3.2.

We know that the sequence $\{f(x, y, \lambda_k)\}$, with the λ_k generated as in (a)–(c), converges to the optimal control. However, it remains to show how to compute $f(x, y, \lambda)$, and hence $\nabla g(\lambda)$ (which is needed in the construction (c)) for a given λ. We take up this problem in the next section.

4. Computing $f(x, y, \lambda)$. For a given λ, $f(x, y, \lambda)$ minimizes the functional (3.1) over the class $\mathscr{C}$, and subject to (2.1). A sequence $\{f_n\}$ converging to $f(x, y, \lambda)$ can be constructed by the following iterative procedure.

(a) Let $f_1 \in \mathscr{C}$ be chosen arbitrarily.

(b) If $f_1, \cdots, f_k \in \mathscr{C}$ have been chosen, choose $\tilde{f}_k \in \mathscr{C}$ to minimize the functional

$$\int\int_\Omega \Big\{ 2(pu_k u + qu_{kx}u_x + qu_{ky}u_y) + rf^2 - \sum_{i=1}^{n} \lambda_i (a_{i0}u + a_{i1}u_x + a_{i2}u_y + b_i f) \Big\}\, dx\, dy.$$

This can be accomplished as follows. First solve

$$v_{xx} + v_{yy} - v = \sum_{i=1}^{n} \lambda_i \left(a_{i0} - \frac{\partial a_{i1}}{\partial x} - \frac{\partial a_{i2}}{\partial y} \right) - 2\{(p - q)u_k - q_x u_{kx} - q_y u_{ky} - q f_k\}$$

on Ω with $v = 0$ on $\partial\Omega$. Then choose $\tilde{f}_k$ by the requirement that the expression

$$v(x, y)f + \sum_{i=1}^{n} \lambda_i b_i(x, y) f - r(x, y) f^2$$

attain its maximum value on I at $f = \tilde{f}_k(x, y)$ for almost all $(x, y) \in \Omega$. It is trivial to verify that $\tilde{f}_k$ so determined has the desired property.

(c) Let $f_{k+1} = f_k + \alpha_k(\tilde{f}_k - f_k)$, where

$$\alpha_k = \frac{\int\int_\Omega r(\Delta f)^2\, dx\, dy}{\int\int_\Omega p(\Delta u)^2 + q(\Delta u_x)^2 + q(\Delta u_y)^2 + r(\Delta f)^2\, dx\, dy}.$$

Here $\Delta f = \tilde{f}_k - f_k$ and $\Delta u = \tilde{u}_k - u_k$.

THEOREM 4.1. *Consider the functional* (3.1) *which assumes its minimum value* m *on* $\mathscr{C}$ *at* $f = f(x, y, \lambda)$. *If we denote the value of this functional at* $f = f_k$ (*from* (a)–(c), § 4) *by* m_k, *then for* $k \geqq 1$,

$$\frac{1}{2}\int\int_\Omega r(f_{k+1}(x, y) - f(x, y, \lambda))^2\, dx\, dy \leqq m_{k+1} - m \leqq (1 - \alpha_k)(m_k - m). \tag{4.1}$$

Proof. Consider the space E introduced in § 2, and the set $K \subset E$. Let $Z_k = (u_k, u_{kx}, u_{ky}, f_k)$, $Z_\lambda = (u(x, y, \lambda), u_x(x, y, \lambda), u_y(x, y, \lambda), f(x, y, \lambda))$ be points in K corresponding to f_k and $f(x, y, \lambda)$. We saw in § 3 that Z_λ is the point in K nearest a given point W_λ in E. By the parallelogram law we have

$$\|Z_\lambda - Z_k\|^2 = 2\|Z_\lambda - W_\lambda\|^2 + 2\|Z_k - W_\lambda\|^2 - 4\|\tfrac{1}{2}(Z_k + Z_\lambda) - W_\lambda\|^2.$$

Since K is convex, $\frac{1}{2}(Z_k + Z_\lambda) \in K$, and so $\|\frac{1}{2}(Z_k + Z_\lambda) - W_\lambda\|^2 \geqq \|Z_\lambda - W_\lambda\|^2$. It follows that

$$\|Z_\lambda - Z_k\|^2 \leqq 2(\|Z_k - W_\lambda\|^2 - \|Z_\lambda - W_\lambda\|^2) = 2(m_k - m).$$

The first inequality in (4.1) follows from this inequality. We shall not prove the second inequality here. Its proof can be constructed, following the method of the proof of convergence given in the first part of the paper [8]. Let us observe, however, that the α_k are bounded away from zero. This follows from (1.8) applied to Δu.

Remark. In carrying out the construction of the sequence $\{f(x, y, \lambda_k)\}$ of Theorem 3.2, one should use $f(x, y, \lambda_k)$ as the initial approximation f_1 to $f(x, y, \lambda_{k+1})$ in computing $f(x, y, \lambda_{k+1})$ according to (a)–(c) of § 4. This provides a good initial approximation for large k. See [9].

Remark. We have said nothing about how best to select s_k in (c) of § 3. The simplest choice is to take $s_k = \nabla g(\lambda_k)$. We have used this choice of s_k in carrying out some experiments in a class of fixed endpoint problems in the control of lumped parameter systems. This work will be reported in [9]. Other choices for s_k will be discussed elsewhere.

5. Inequality constraints. In this section we indicate how the results developed so far can be applied to problems with linear inequality constraints. If some or all of the constraints (2.3) were replaced by inequality constraints of the form $a_i \leqq c_i$, a_i defined by (2.4), then the corresponding multipliers in Theorem 2.2 would be $\leqq 0$. In this case the duality theory we have discussed is still valid, except that now, the maximization (3.2) must be taken subject to $\lambda_i \leqq 0$ if λ_i corresponds to an inequality constraint. The gradient scheme (a)–(c) of § 3 can be modified to solve this constrained optimization problem. All that is required is that we do gradient projection. This means that in place of (c) of § 3, we choose λ_{k+1} according to the rule $\lambda_{k+1} = \min\{0, \lambda_k + \theta t_k s_k\}$, where this equation is defined componentwise, for the components corresponding to inequality constraints. All other components should be computed as in (c), § 3.

Often we encounter variational problems with constraints that cannot be expressed in the form (2.4) with equality or inequality. The constraint $u \geqq 0$ imposed on the surface in Fig. 1 is one such example. We wish to discuss this type of constraint briefly here. For simplicity we consider only the example

$$\text{(5.1)} \qquad \text{minimize} \quad c_0(u) = \int\int_\Omega u^2 + u_x^2 + u_y^2 \, dx \, dy$$

subject to

$$u = \varphi \quad \text{on } \partial\Omega, \qquad u \geqq 0 \quad \text{on } \Omega.$$

φ and Ω are as in § 4. This problem is discussed in [10, p. 29] and in [11]. It is also clearly closely related to the example we discussed in the Introduction. And in fact, the present discussion can be combined with what we did earlier to provide a method for minimizing the functional (1.4) subject to the constraints (1.3) and

subject to $u \geqq 0$. This provides us with a solution to a problem which more closely approximates the example we described in the Introduction.

Define the function g on the class of square integrable functions on Ω by

$$g(v) = \min_{u} \int\int_{\Omega} -v(x, y)u(x, y) + u^2 + u_x^2 + u_y^2 \, dx \, dy, \tag{5.2}$$

where the minimization is taken over the class of functions u, which are square integrable on Ω and satisfy $u = \varphi$ on $\partial\Omega$. It is easy to show that the problem

$$\max_{v \geqq 0} g(v) \tag{5.3}$$

is dual to the problem (5.1).

The directional derivative of g at v, in the direction $h(x, y)$, is given by

$$\delta g(v, h) = \int\int_{\Omega} -u_v h \, dx \, dy. \tag{5.4}$$

u_v is the function which accomplishes the minimization in (5.2). The technique for establishing (5.4) is demonstrated in the proof of Theorem 2.1 in [4]. The gradient projection method for solving (5.3) will now be described. The details of the convergence proof are left to the reader. A good discussion of the gradient projection method is contained in [12].

(a) Let θ and β be as in (a), § 3.

(b) Let $v_1 \geqq 0$ be chosen arbitrarily in $L^2(\Omega)$.

(c) If $v_1, \cdots, v_k \geqq 0$ have been chosen in $L_2(\Omega)$, take $v_{k+1}(x, y) = \max\{0, v_k(x, y) + \theta\sigma_k S(x, y)\}$, where S is any function in $L_2(\Omega)$ satisfying

$$-u_{v_k}(x, y)S(x, y) \geqq u_{v_k}^2(x, y),$$

$$S^2(x, y) \leqq \beta u_{v_k}^2(x, y) \quad \text{for almost all } (x, y) \in \Omega,$$

$$\sigma_k = \frac{\int\int_{\Omega} u_{v_k}^2(x, y) \, dx \, dy}{\int\int_{\Omega} S^2(x, y) \, dx \, dy}.$$

REFERENCES

[1] G. Hellwig, *Partial Differential Equations*, Blaisdell, New York, 1964.

[2] E. B. Lee, *Linear optimal control problems with isoperimetric constraints*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 87–90.

[3] S. Saks, *Theory of the Integral*, Hafner, New York, 1937.

[4] B. N. Pshenichniy, *Linear optimal control problems*, this Journal, 4 (1966), pp. 577–593.

[5] J. S. Meditch and L. W. Neustadt, *An application of optimal control to midcourse guidance*, Proc. Second IFAC Congress, Butterworths, London, 1965, pp. 292–299.

[6] L. W. Neustadt and B. H. Paiewonsky, *On synthesizing optimal controls*, Ibid., pp. 283–291.

[7] J. L. Kelly and Isaac Namioka, *Linear Topological Spaces*, Van Nostrand, Princeton, N.J., 1963.

[8] E. R. Barnes, *Computing optimal controls in systems with distributed parameters*, Proc. IFAC Symposium on Optimal Control of Distributed Parameter Systems, Banff, Alberta, June 1971.

[9] E. R. BARNES, *Solution of a dual problem in optimal control theory, I*, this Journal, submitted.
[10] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux derivées partielles*, Dunod-Gauthier Villars, Paris, 1968; English transl., Springer-Verlag, New York, 1970.
[11] H. R. BREZIS AND G. STAMPACCHIA, *Sur la régularité de la solution d'inéquations élliptiques*, Bull. Soc. Math. France, 96 (1968), pp. 153–180.
[12] E. S. LEVITIN AND B. T. POLYAK, *Constrained minimization methods*, Zh. Vychisl. Mat. i Mat. Fiz., 6 (1966), no. 5, pp. 787–823 = U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), no. 5, pp. 1–50.

SIAM J. CONTROL
Vol. 10, No. 2, May 1972

COMPARISON OF DIFFERENTIAL GAMES OF FIXED DURATION*

A. N. V. RAO†

Abstract. The upper and lower Values of two differential games, one described by a system of n differential equations and the other, by a system of m differential equations, are compared. The approach taken depends upon the use of Lyapunov-like functions and the theory of differential inequalities. It is suggested that such comparison results are of interest because they enable us to obtain information about a given game by studying a simpler game.

1. Introduction. The purpose of this paper is to compare the Values of two differential games of fixed duration, one described by a system of n differential equations and the other by a system of m differential equations. Such comparison results are of interest because they enable us to obtain information about a given game by studying a simpler game. More specifically, we wish to obtain an estimate of the Value of a given game in terms of the Value of a comparison game, which is, in general, simpler. The comparison principle, which has been a powerful tool in studying the qualitative behavior of differential equations, is appropriately extended to yield such estimates. Our approach therefore depends upon the systematic use of Lyapunov-like functions and the theory of differential inequalities.

Consider the differential game, whose state $x(t)$ is determined by the differential system

$$\dot{x}(t) = f_1(t, x(t), y(t), z(t)), \qquad x(t_0) = x_0, \tag{1.1}$$

where $f_1 \in C[[t_0, T] \times R^n \times Y \times Z, R^n]$ ($f \in C[E_1, E_2]$ will mean that f is a continuous map from E_1 into E_2) and $y(t)$ and $z(t)$ are measurable functions of t, with values in Y and Z, for almost all $t \in [t_0, T]$. It will be assumed that Y and Z are fixed compact subsets of R^{p_1} and R^{p_2}, respectively. The payoff associated with (1.1) is a given real-valued functional

$$P_1 = \mu(x) + \int_{t_0}^{T} h_1(s, x(s), y(s), z(s))\, ds, \tag{1.2}$$

where μ is a real functional on $C[[t_0, T], R^n]$ and h_1 is a real function on $[t_0, T] \times R^n \times Y \times Z$.

Let the state $u(t)$ of the comparison game be determined by the differential system

$$\dot{u}(t) = f_2(t, u(t), v(t), w(t)), \qquad u(t_0) = u_0, \tag{1.3}$$

where $f_2 \in C[[t_0, T] \times R^m \times V \times W, R^m]$ and $v(t)$ and $w(t)$ are measurable functions of t, with values in V and W, for almost all $t \in [t_0, T]$ (V and W are some

* Received by the editors May 21, 1971, and in revised form October 22, 1971. Presented at the NSF Regional Conference on Control Theory, held at the University of Maryland Baltimore County, August 23–27, 1971.

† Department of Mathematics, University of Rhode Island, Kingston, Rhode Island. Currently at Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

fixed compact subsets of R^{q_1} and R^{q_2}, respectively). Let the payoff be given by

$$P_2 = \lambda(u) + \int_{t_0}^{T} h_2(s, u(s), v(s), w(s))\, ds, \tag{1.4}$$

where λ is a real functional on $C[[t_0, T], R^m]$ and h_2 is a real function on $[t_0, T] \times R^m \times V \times W$.

Both the games start at time $t = t_0$ and end at time $t = T$.

2. Preliminaries. Consider the games (1.1), (1.2) and (1.3), (1.4). The measurable functions $y(t)$, $z(t)$, $v(t)$ and $w(t)$ will be called control functions. In game (1.1), (1.2), a player, whom we shall call y, regulates the control $y(t)$ so as to maximize the payoff P_1, while a second player, called z, regulates the control $z(t)$ so as to minimize P_1. Similarly, in the game (1.3), (1.4), v (with control $v(t)$) tries to maximize P_2, while w (with control $w(t)$) tries to minimize it.

Let $\mathscr{C}[t_0, T]$ denote the Banach space of complex-valued continuous functions on $[t_0, T]$, with the sup norm. Let $[\mathscr{C}[t_0, T]]^n$ denote the product of n spaces $\mathscr{C}[t_0, T]$. Let $X_{t_0,T}$ denote the subset of $[\mathscr{C}[t_0, T]]^n$ consisting of all trajectories of (1.1). $U_{t_0,T}$ is similarly defined with respect to (1.3).

The concept of δ-games, upper and lower δ-strategies, upper and lower δ-Values and upper and lower Values are all defined as in [1]. We denote upper and lower δ-strategies for y, in game (1.1), (1.2), by $\Gamma^\delta, \Gamma_\delta$. For z, these will be denoted by $\Delta^\delta, \Delta_\delta$. The controls corresponding to these strategies will be denoted, respectively, by $y^\delta(t), y_\delta(t), z^\delta(t)$ and $z_\delta(t)$. A trajectory of (1.1) in an upper δ-game will be denoted by $x^\delta(t)$ and in a lower δ-game by $x_\delta(t)$.

Upper and lower δ-strategies for v and w, in game (1.3), (1.4), will be denoted by $\Sigma^\delta, \Sigma_\delta, \Lambda^\delta$ and Λ_δ, respectively; the corresponding controls by $v^\delta(t), v_\delta(t), w^\delta(t)$ and $w_\delta(t)$. Trajectories of (1.3) in upper and lower δ-games will be denoted, respectively, by $u^\delta(t)$ and $u_\delta(t)$.

We shall denote the upper and lower δ-Values for the game (1.1), (1.2) by V_1^δ and $V_{1,\delta}$; the upper and lower Values by V_1^+ and V_1^-. For the game (1.3), (1.4), these will be denoted, respectively, by $V_2^\delta, V_{2,\delta}, V_2^+$ and V_2^-. If Γ^δ and Δ_δ are two strategies in game (1.1), (1.2), the corresponding payoff will be denoted by $P_1[\Delta_\delta, \Gamma^\delta]$. $P_2[\Lambda_\delta, \Sigma^\delta]$ is similarly defined.

We shall need the following concept of quasi-monotonicity of a function [2]. Let $g(t, u) \in C[[t_0, T] \times R^n, R^n]$.

DEFINITION. The function $g(t, u)$ is said to be *quasi-monotone nondecreasing* if its pth component, $g_p(t, u)$, is nondecreasing in $u_j, j = 1, 2, \cdots, n$ and $j \neq p$; $p = 1, 2, \cdots, n$.

We shall make the following assumptions with respect to (1.1) and (1.3).

(H_1)(i) For each pair of control functions $y(t)$ and $z(t)$, (1.1) has a unique solution.

(ii) The solutions of (1.1) are uniformly bounded on $[t_0, T]$.

(H_2)(i) For each pair of control functions $v(t)$ and $w(t)$, (1.3) has a unique solution.

(ii) The solutions of (1.3) are uniformly bounded on $[t_0, T]$.

It is known that under the assumptions (H_1) and (H_2), the games (1.1), (1.2) and (1.3), (1.4) have both upper and lower Values [1].

In order to connect the two differential games that we are considering, we shall need the following assumptions:

(H_3) There exist functions $L \in C[[t_0, T] \times R^n, R^m]$, $L_1 \in C[Y, V]$, $L_2 : Z \to W$, with $L(t, x)$ locally Lipschitzian in x and L_2 onto, such that

$$D^+L(t,x) \equiv \limsup_{h \to 0^+} \frac{1}{h}[L(t+h, x+hf_1(t,x,y,z)) - L(t,x)]$$

$$\leqq f_2(t, L(t,x), L_1(y), L_2(z)),$$

for $(t, x, y, z) \in [t_0, T] \times R^n \times Y \times Z$.

(H_4) For $(t, x, y, z) \in [t_0, T] \times R^n \times Y \times Z$, the functions h_1 and h_2 satisfy the inequality $h_1(t, x, y, z) \leqq h_2(t, L(t,x), L_1(y), L_2(z))$.

(H_5) For $\phi \in C[[t_0, T], R^n]$, $\mu(\phi) \leqq \lambda(L[\phi])$, where $L[\phi] \in C[[t_0, T], R^m]$ and is such that $L[\phi](t) = L(t, \phi(t))$.

All the vector inequalities used above and subsequently are to be understood as componentwise. Also, if $u_1, u_2 \in C[[t_0, T], R^m]$, we shall say that $u_1 \leqq u_2$, if $u_1(t) \leqq u_2(t)$ for all $t \in [t_0, T]$.

3. Main results. Now we are in a position to state our main results.

THEOREM 1. *Let the games* (1.1), (1.2) *and* (1.3), (1.4) *satisfy the assumptions* (H_1)–(H_5). *Further, assume that*

(a) *on some nonempty set* $Z_0 \subset Z$, *the function* L_2 *has a continuous inverse*;

(b) $f_2(t, u, v, w)$ *is quasi-monotone nondecreasing in* u;

(c) $h_2(t, u, v, w)$ *and* $\lambda(u)$ *are monotone nondecreasing in* u;

(d) $L(t_0, x_0) \leqq u_0$.

Then (i) $V_1^+ \leqq V_2^+$; (ii) $V_1^- \leqq V_2^-$.

Proof. We shall prove (i). Let us consider an upper δ-game. In view of the assumptions (H_1) and (H_2) and the definition of an upper δ-Value, it follows that, for each $\varepsilon > 0$, there is a lower δ-strategy $\tilde{\Lambda}_\delta$ for w and an upper δ-strategy $\tilde{\Gamma}^\delta$ for y, such that

$$V_2^\delta \geqq P_2[\tilde{\Lambda}_\delta, \Sigma^\delta] - \varepsilon \quad \text{for all } \Sigma^\delta \tag{3.1}$$

and

$$V_1^\delta \leqq P_1[\Delta_\delta, \tilde{\Gamma}^\delta] + \varepsilon \quad \text{for all } \Delta_\delta. \tag{3.2}$$

We shall show that there exist strategies $\Delta_\delta, \Sigma^\delta$ such that the controls z_δ, y^δ corresponding to $\Delta_\delta, \tilde{\Gamma}^\delta$ and the controls w_δ, v^δ corresponding to $\tilde{\Lambda}_\delta, \Sigma^\delta$ satisfy:

$$v^\delta(t) = L_1(y^\delta(t)) \quad \text{and} \quad w_\delta(t) = L_2(z_\delta(t)). \tag{3.3}$$

For this, let w choose $w_{1,\delta} = \tilde{\Lambda}_{\delta,1}$ on the first interval I_1 of the δ-game. Then, let z choose $z_{1,\delta}$ such that $z_{1,\delta}(t) \in Z_0$ and $L_2(z_{1,\delta}(t)) = w_{1,\delta}(t)$, on I_1. It is clear from the assumptions on L_2 that such a choice for z is not only possible but also unique. Using the strategy $\tilde{\Gamma}^\delta$, let y choose $y_{1,\delta} = \tilde{\Gamma}^{\delta,1} z_{\delta,1}$ on I_1. Now, let v choose $v^\delta(t) = L_1(y^\delta(t))$, $t \in I_1$. Again, the choice of v on I_1 is uniquely defined. Using the strategy $\tilde{\Lambda}_\delta$, w chooses $w_{2,\delta}$ on I_2 such that $w_{2,\delta} = \tilde{\Lambda}_{\delta,1} v_{\delta,1}$ on I_2 and the scheme is repeated. Let the controls thus obtained for y, z, v and w be respectively denoted $y^\delta(t), z_\delta(t), v^\delta(t)$ and $w_\delta(t)$ for $t \in [t_0, T]$.

Since on each of the intervals I_j, $j = 2, 3, \cdots, N$, the choice of z-control is uniquely dictated by the choices of y-controls on the intervals $I_1, I_2, \cdots, I_{j-1}$, it follows that there exists a lower δ-strategy, Δ_δ, for z, corresponding to the control z_δ. Similarly, there is an upper δ-strategy, Σ^δ, for v, corresponding to the control $v^\delta(t)$. Further, it is obvious from the choices of the controls, that they verify (3.3).

Let the trajectories of (1.1) and (1.3), corresponding to the above controls, be denoted by $x^\delta(t)$ and $u^\delta(t)$, respectively. Then, from the assumption (H_4), we have

$$h_1(t, x^\delta(t), y^\delta(t), z_\delta(t)) \leqq h_2(t, L(t, x^\delta(t)), L_1(y^\delta(t)), L_2(z_\delta(t))). \tag{3.4}$$

Furthermore, from a result on differential inequalities [2], it follows that

$$L(t, x^\delta(t)) \leqq u^\delta(t). \tag{3.5}$$

(Note that by assumption f_2 is Lipschitzian; hence the maximal solution of (1.3) is the unique solution $u^\delta(t)$.) Since $h_2(t, u, v, w)$ is nondecreasing in u, we conclude from (3.4) and (3.5) that

$$h_1(t, x^\delta(t), y^\delta(t), z_\delta(t)) \leqq h_2(t, u^\delta(t), v^\delta(t), w_\delta(t)). \tag{3.6}$$

Also, from (H_5) and (3.5), we obtain

$$\mu(x^\delta) \leqq \lambda(u^\delta). \tag{3.7}$$

The inequalities (3.6) and (3.7) allow us to conclude that

$$\begin{aligned} P_1[\Delta_\delta, \tilde{\Gamma}^\delta] &\equiv \mu(x^\delta) + \int_{t_0}^{T} h_1(s, x^\delta(s), y^\delta(s), z_\delta(s))\, ds \\ &\leqq \lambda(u^\delta) + \int_{t_0}^{T} h_2(s, u^\delta(s), v^\delta(s), w_\delta(s))\, ds \equiv P_2[\tilde{\Lambda}_\delta, \Sigma^\delta]. \end{aligned} \tag{3.8}$$

The assertion of the theorem now follows from the inequalities (3.1), (3.2) and (3.8), on taking the limit as $\varepsilon, \delta \to 0$.

THEOREM 2. *Let the assumptions of Theorem* 1 *hold, except that* (H_1)(i) *and* (H_2)(i) *are replaced by the assumptions that* f_1 *and* f_2 *are locally Lipschitzian in* x *and* u, *respectively. Let* h_1, h_2 *be continuous, let* μ, λ *be uniformly continuous and let the following decomposition hold*:

$$\begin{aligned} f_1(t, x, y, z) &= f_{1,1}(t, x, y) + f_{1,2}(t, x, z); \\ f_2(t, u, v, w) &= f_{2,1}(t, u, v) + f_{2,2}(t, u, w); \\ h_1(t, x, y, z) &= h_{1,1}(t, x, y) + h_{1,2}(t, x, z); \\ h_2(t, u, v, w) &= h_{2,1}(t, u, v) + h_{2,2}(t, u, w). \end{aligned}$$

Then $V_1 \leqq V_2$.

Proof. It is known [1], that if the above assumptions hold, the games (1.1), (1.2) and (1.3), (1.4) have Values. The result now follows from Theorem 1.

Remark 1. Analogues of Theorems 1 and 2, yielding lower estimates of the Value of the given differential game, can similarly be proved.

Remark 2. If the comparison game is a scalar game, then the assumption of quasi-monotonicity of f_2 is redundant, as, in this case, the assumption is always satisfied.

Example. Let us give an example to illustrate the theorem. Consider the differential game determined by

$$\begin{aligned} \dot{x}_1 &= x_2 - a_1 x_1^3 + y + z, \\ \dot{x}_2 &= -x_1 - a_2 x_2^3 + y - z, \quad x_1(0) = 0, \quad x_2(0) = \sqrt{2}, \quad a_1, a_2 \geqq 0. \end{aligned} \tag{3.9}$$

Let the payoff be

$$P_1 = \int_0^1 [\tfrac{1}{2} x_1^2(s) + \tfrac{1}{2} x_2^2(s)]\, ds. \tag{3.10}$$

The controls y and z are constrained by the condition $-1 \leqq y, z \leqq 1$. Choose the comparison game to be

$$\dot{u} = u + v + w, \quad u(0) = 1, \quad 0 \leqq v, w \leqq 1,$$

with the payoff

$$P_2 = \int_0^1 u(s)\, ds. \tag{3.11}$$

Define (i) $L(t, x) = x_1^2/2 + x_2^2/2$, (ii) $L_1 : L_1 y = y^2 = v$, (iii) $L_2 : L_2 z = z^2 = w$, (iv) $Z_0 = [0, 1]$. With these choices, it is easy to verify that the assumptions of Theorem 2 are satisfied (it is well known that (H_1) holds if $f(t, x, y, z)$ is continuous, Lipschitz continuous in x in bounded sets and

$$x \cdot f(t, x, y, z) \leqq c_1 |x|^2 + c_2 \qquad (c_1 \text{ and } c_2 \text{ constants})$$

for all (t, x, y, z)). Hence the Value of the given game is dominated by the Value of the comparison game. The Value of the comparison game is easily computed to be $2e - 3$. Therefore $V_1 \leqq 2e - 3$.

Acknowledgments. The author is deeply indebted to Dr. Roxin and Dr. Lakshmikantham, for extensive creative discussions and comments. Thanks are also due to the referee, for his useful suggestions.

REFERENCES

[1] A. FRIEDMAN, *On the definition of differential games and the existence of Value and saddle points*, J. Differential Equations, 7 (1970), pp. 69–91.

[2] V. LAKSHMIKANTHAM AND S. LEELA, *Differential and Integral Inequalities, Theory and Applications*, vol. 1, Academic Press, New York, 1969.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

THE SINGULARLY PERTURBED LINEAR STATE REGULATOR PROBLEM*

R. E. O'MALLEY, JR.†

Abstract. We shall seek the optimal control and corresponding trajectories for the state regulator problem

$$\frac{dx}{dt} = A_1(t,\varepsilon)x + A_2(t,\varepsilon)z + B_1(t,\varepsilon)u,$$

$$\varepsilon\frac{dz}{dt} = A_3(t,\varepsilon)x + A_4(t,\varepsilon)z + B_2(t,\varepsilon)u$$

on the interval $0 \leqq t \leqq 1$ with the initial state vector $y(0,\varepsilon)$ prescribed $\left(\text{for } y = \begin{pmatrix} x \\ z \end{pmatrix}\right)$ and where the quadratic cost functional

$$J(\varepsilon) = \frac{1}{2}y'(1,\varepsilon)\pi(\varepsilon)y(1,\varepsilon) + \frac{1}{2}\int_0^1 [y'(t,\varepsilon)Q(t,\varepsilon)y(t,\varepsilon) + u'(t,\varepsilon)R(t,\varepsilon)u(t,\varepsilon)]\,dt$$

is minimized. Under appropriate hypotheses, it will be possible to obtain a complete asymptotic solution uniformly valid as $\varepsilon \to 0$. Such results are of value in practical control theory where ε represents certain often neglected "parasitic" parameters whose presence causes the order of the model to increase.

1. Problem statement and method of solution. We shall consider the linear state regulator problems consisting of the system

$$\begin{aligned} \frac{dx}{dt} &= A_1(t,\varepsilon)x + A_2(t,\varepsilon)z + B_1(t,\varepsilon)u, \\ \varepsilon\frac{dz}{dt} &= A_3(t,\varepsilon)x + A_4(t,\varepsilon)z + B_2(t,\varepsilon)u \end{aligned} \tag{1.1}$$

on the interval $0 \leqq t \leqq 1$ (or, equivalently, any closed, bounded interval), the initial conditions

$$\begin{aligned} x(0,\varepsilon) &= x^0(\varepsilon), \\ z(0,\varepsilon) &= z^0(\varepsilon) \end{aligned} \tag{1.2}$$

and the quadratic cost functional

$$\begin{aligned} J(\varepsilon) &= \tfrac{1}{2}y'(1,\varepsilon)\pi(\varepsilon)y(1,\varepsilon) \\ &\quad + \frac{1}{2}\int_0^1 [y'(t,\varepsilon)Q(t,\varepsilon)y(t,\varepsilon) + u'(t,\varepsilon)R(t,\varepsilon)u(t,\varepsilon)]\,dt, \end{aligned} \tag{1.3}$$

* Received by the editors April 29, 1971.

† Division of Electromagnetic Research, Courant Institute of Mathematical Sciences, and Department of Mathematics, New York University, University Heights, New York, New York 10012. This research was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-71-2013.

where the prime denotes transposition,

$$y = \begin{bmatrix} x \\ z \end{bmatrix}, \quad Q = \begin{bmatrix} Q_1 & Q_2 \\ Q_2' & Q_3 \end{bmatrix} \quad \text{and} \quad \pi = \begin{bmatrix} \pi_1 & \varepsilon\pi_2 \\ \varepsilon\pi_2' & \varepsilon\pi_3 \end{bmatrix}$$

for n-vector x, m-vector z, and r-vector u. When $\varepsilon > 0$ is fixed, it is well known that the elementary control problem has, under appropriate hypotheses, a unique optimal control which minimizes J (see, e.g., Kalman, Falb and Arbib [6]). We are interested, however, in obtaining the asymptotic solution of the problem as the small positive parameter ε tends to zero. Such singular perturbation problems are of considerable significance in practical situations where ε represents certain often neglected "parasitic" parameters whose presence causes the order of the mathematical model to increase (cf. Sannuti and Kokotović [14]). Hadlock, Jamshidi and Kokotović [2] give an example of a sixth order model for the optimal tension regulation of a strip winding process for a rolling mill plant. They show that asymptotic results are far superior to the physically unacceptable results obtained when $\varepsilon = 0$.

We shall assume that the matrices $A_i(t, \varepsilon)$, $B_i(t, \varepsilon)$, $x^0(\varepsilon)$, $z^0(\varepsilon)$, $\pi_i(\varepsilon)$, $Q_i(t, \varepsilon)$ and $R(t, \varepsilon)$ all have asymptotic power series expansions as $\varepsilon \to 0$ uniformly in $0 \leqq t \leqq 1$, that the $A_i(t, \varepsilon)$, $B_i(t, \varepsilon)$, $Q_i(t, \varepsilon)$, and $R(t, \varepsilon)$ are infinitely differentiable functions of t in $0 \leqq t \leqq 1$, and that the matrix $R(t, \varepsilon)$ is symmetric and positive definite while $Q(t, \varepsilon)$ and $\pi(\varepsilon)$ are symmetric and positive semidefinite throughout $0 \leqq t \leqq 1$ for $\varepsilon \geqq 0$ sufficiently small.

To obtain necessary and sufficient conditions for an optimum control, we introduce the Hamiltonian

$$\begin{aligned} H(x, z, u, p_1, p_2, t, \varepsilon) = {} & \tfrac{1}{2}(x'Q_1x + 2x'Q_2z + z'Q_3z + u'Ru) \\ & + p_1'(A_1x + A_2z + B_1u) + p_2'(A_3x + A_4z + B_2u). \end{aligned} \tag{1.4}$$

Elementary calculus of variations (cf. Kalman, Falb and Arbib) implies that along an optimal trajectory,

$$\frac{\partial H}{\partial u} = Ru + B_1'p_1 + B_2'p_2 = 0, \tag{1.5}$$

where the costate vectors $p_1(t, \varepsilon)$ and $p_2(t, \varepsilon)$ satisfy the equations

$$\begin{aligned} \frac{dp_1}{dt} &= \frac{-\partial H}{\partial x} = -Q_1x - Q_2z - A_1'p_1 - A_3'p_2, \\ \varepsilon\frac{dp_2}{dt} &= \frac{-\partial H}{\partial z} = -Q_2'x - Q_3z - A_2'p_1 - A_4'p_2 \end{aligned} \tag{1.6}$$

on $0 \leqq t \leqq 1$ and the terminal conditions

$$\begin{aligned} p_1(1, \varepsilon) &= \pi_1(\varepsilon)x(1, \varepsilon) + \varepsilon\pi_2(\varepsilon)z(1, \varepsilon), \\ p_2(1, \varepsilon) &= \pi_2'(\varepsilon)x(1, \varepsilon) + \pi_3(\varepsilon)z(1, \varepsilon). \end{aligned} \tag{1.7}$$

Together with (1.5)–(1.7), we have the original state equations

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial H}{\partial p_1} = A_1x + A_2z + B_1u, \\ \varepsilon\frac{dz}{dt} &= \frac{\partial H}{\partial p_2} = A_3x + A_4z + B_2u \end{aligned} \tag{1.8}$$

with the initial conditions

$$\begin{aligned} x(0,\varepsilon) &= x^0(\varepsilon), \\ z(0,\varepsilon) &= z^0(\varepsilon). \end{aligned} \tag{1.9}$$

Note that since $\partial^2 H/\partial u^2 = R$ is positive definite, the optimal control will minimize $J(\varepsilon)$.

From (1.5), we have the control law

$$u(t,\varepsilon) = -R^{-1}(t,\varepsilon)(B_1'(t,\varepsilon)p_1(t,\varepsilon) + B_2'(t,\varepsilon)p_2(t,\varepsilon)), \tag{1.10}$$

and there remains the linear system

$$\begin{aligned} \frac{dx}{dt} &= A_1 x + A_2 z - S_1 p_1 - S p_2, \\ \frac{dp_1}{dt} &= -Q_1 x - Q_2 z - A_1' p_1 - A_3' p_2, \\ \varepsilon\frac{dz}{dt} &= A_3 x + A_4 z - S' p_1 - S_2 p_2, \\ \varepsilon\frac{dp_2}{dt} &= -Q_2' x - Q_3 z - A_2' p_1 - A_4' p_2, \end{aligned} \tag{1.11}$$

while S_1, S and S_2 are the matrices

$$\begin{aligned} S_1(t,\varepsilon) &= B_1(t,\varepsilon)R^{-1}(t,\varepsilon)B_1'(t,\varepsilon), \\ S(t,\varepsilon) &= B_1(t,\varepsilon)R^{-1}(t,\varepsilon)B_2'(t,\varepsilon), \\ S_2(t,\varepsilon) &= B_2(t,\varepsilon)R^{-1}(t,\varepsilon)B_2'(t,\varepsilon). \end{aligned}$$

We must solve the system (1.11) subject to the $2n + 2m$ boundary conditions given by (1.7) and (1.9). Since the order of this system drops from $2n + 2m$ for $\varepsilon > 0$ to $2n$ for $\varepsilon = 0$, we have a singularly perturbed two-point boundary value problem.

By analysis of similar problems (see, e.g., Harris [3] or O'Malley [10]) or by examining solutions of constant coefficient problems, one is lead to seek asymptotic solutions of the form

$$\begin{aligned} x(t,\varepsilon) &= X(t,\varepsilon) + \varepsilon m_1(\kappa,\varepsilon) + \varepsilon n_1(\sigma,\varepsilon), \\ z(t,\varepsilon) &= Z(t,\varepsilon) + m_2(\kappa,\varepsilon) + n_2(\sigma,\varepsilon), \\ p_1(t,\varepsilon) &= P_1(t,\varepsilon) + \varepsilon\rho_1(\kappa,\varepsilon) + \varepsilon\gamma_1(\sigma,\varepsilon), \\ p_2(t,\varepsilon) &= P_2(t,\varepsilon) + \rho_2(\kappa,\varepsilon) + \gamma_2(\sigma,\varepsilon). \end{aligned} \tag{1.12}$$

Here, the "outer expansion"

$$(X(t,\varepsilon), Z(t,\varepsilon), P_1(t,\varepsilon), P_2(t,\varepsilon)) \tag{1.13}$$

will satisfy the system (1.11) throughout $0 \leqq t \leqq 1$ and will have an asymptotic power series expansion there as $\varepsilon \to 0$. It will not, in general, satisfy the boundary conditions (1.7) and (1.9), however. The expansion

$$(\varepsilon m_1(\kappa,\varepsilon), m_2(\kappa,\varepsilon), \varepsilon\rho_1(\kappa,\varepsilon), \rho_2(\kappa,\varepsilon)) \tag{1.14}$$

represents the "boundary layer correction" at $t = 0$. It will have an asymptotic power series expansion as $\varepsilon \to 0$ whose terms tend to zero as the left boundary layer coordinate

$$\kappa = t/\varepsilon \tag{1.15}$$

tends to infinity. Likewise,

$$(\varepsilon n_1(\sigma, \varepsilon), n_2(\sigma, \varepsilon), \varepsilon\gamma_1(\sigma, \varepsilon), \gamma_2(\sigma, \varepsilon)) \tag{1.16}$$

represents the "boundary layer correction" at $t = 1$. The terms of its asymptotic series will tend to zero as the right boundary layer coordinate

$$\sigma = (1 - t)/\varepsilon \tag{1.17}$$

tends to infinity. Thus, (1.12) implies that the solution (x, z, p_1, p_2) tends to the outer expansion (X, Z, P_1, P_2) within $(0, 1)$ to all orders ε^j as $\varepsilon \to 0$. Convergence will not be uniform as $\varepsilon \to 0$ at $t = 0$ and $t = 1$, however, where derivatives of the solution will become unbounded as $\varepsilon \to 0$. (Note that similar boundary layer methods for initial value problems are discussed in Vasil'eva [17] and O'Malley [11].)

By substituting the outer expansion (1.13) into the system (1.11) and setting $\varepsilon = 0$, the leading term of the outer expansion must satisfy the reduced system

$$\begin{aligned}
\frac{dX_0}{dt} &= A_{10}X_0 + A_{20}Z_0 - S_{10}P_{10} - S_0P_{20}, \\
\frac{dP_{10}}{dt} &= -Q_{10}X_0 - Q_{20}Z_0 - A'_{10}P_{10} - A'_{30}P_{20}, \\
0 &= A_{30}X_0 + A_{40}Z_0 - S'_0P_{10} - S_{20}P_{20}, \\
0 &= -Q'_{20}X_0 - Q_{30}Z_0 - A'_{20}P_{10} - A'_{40}P_{20}.
\end{aligned} \tag{1.18}$$

(Here and below, f_j represents the jth term of an asymptotic power series $f(\varepsilon)$ such that $f(\varepsilon) \sim \sum_{j=0}^{\infty} f_j\varepsilon^j$ as $\varepsilon \to 0$.) Note that (1.18) can be solved for Z_0 and P_{20} as linear functions of X_0 and P_{10} provided that the $2m \times 2m$ matrix

$$G(t) = \begin{pmatrix} A_{40} & -S_{20} \\ -Q_{30} & -A'_{40} \end{pmatrix} \tag{1.19}$$

is nonsingular throughout $0 \leqq t \leqq 1$. We shall obtain

$$\begin{aligned}
Z_0 &= \tilde{K}_{10}X_0 + \tilde{K}_{20}P_{10}, \\
P_{20} &= \tilde{K}_{30}X_0 + \tilde{K}_{40}P_{10},
\end{aligned}$$

where the $\tilde{K}_{j0}$'s can be explicitly obtained and (1.18) becomes

$$\begin{aligned}
\frac{dX_0}{dt} &= K_{10}X_0 + K_{20}P_{20}, \\
\frac{dP_{10}}{dt} &= K_{30}X_0 + K_{40}P_{10}
\end{aligned} \tag{1.20}$$

where, e.g., $K_{10} = A_{10} + A_{20}\tilde{K}_{10} - S_0\tilde{K}_{30}$.

We shall assume that the system (1.20) has a unique solution on the interval $0 \leqq t \leqq 1$ which satisfies the $2n$ boundary conditions

$$\begin{aligned} X_0(0) &= x^0(0), \\ P_{10}(1) &= \pi_{10} X_0(1). \end{aligned} \tag{1.21}$$

This will simply require that an appropriate $2n \times 2n$ determinant be nonzero. Thus, the leading term of the outer expansion (1.13), and thereby the limiting solution within $(0, 1)$, will satisfy the reduced problem which consists of the system (1.11) and the boundary conditions for $x(0, \varepsilon)$ and $p_1(1, \varepsilon)$ all evaluated at $\varepsilon = 0$. The boundary conditions prescribed for $z(0, \varepsilon)$ and $P_2(1, \varepsilon)$ are cancelled in defining the reduced problem because it cannot be expected, in general, to satisfy $2m + 2n$ boundary conditions. In particular, we shall generally have $Z_0(0) \neq z^0(0)$ and $P_{20}(1) \neq \pi'_{20} X_0(1) + \pi_{30} Z_0(1)$.

We shall be able to uniquely obtain the complete expansion (1.12) under the following four assumptions.

(H1) The reduced problem

$$\begin{aligned} \frac{dx}{dt} &= A_1(t,0)x + A_2(t,0)z - B_1(t,0)R^{-1}(t,0)B_1'(t,0)p_1 \\ &\quad - B_1(t,0)R^{-1}(t,0)B_2'(t,0)p_2, \\ 0 &= A_3(t,0)x + A_4(t,0)z - B_2(t,0)R^{-1}(t,0)B_1'(t,0)p_1 \\ &\quad - B_2(t,0)R^{-1}(t,0)B_2'(t,0)p_2, \\ \frac{dp_1}{dt} &= -Q_1(t,0)x - Q_2(t,0)z - A_1'(t,0)p_1 - A_3'(t,0)p_2, \\ 0 &= -Q_2'(t,0)x - Q_3(t,0)z - A_2'(t,0)p_1 - A_4'(t,0)p_2 \end{aligned} \tag{1.22}$$

with $x(0) = x_0(0)$ and $p_1(1) = \pi_1(0)x(1)$ has a unique solution.

(H2) The $2m \times 2m$ matrix

$$G(t) = \begin{pmatrix} A_4(t,0) & -B_2(t,0)R^{-1}(t,0)B_2'(t,0) \\ -Q_3(t,0) & -A_4'(t,0) \end{pmatrix} \tag{1.23}$$

has m eigenvalues $\lambda_1(t), \cdots, \lambda_m(t)$ with negative real parts throughout $0 \leqq t \leqq 1$ and m eigenvalues $\lambda_{m+1}(t), \cdots, \lambda_{2m}(t)$ with positive real parts there.

(H3) Using Wasow's definition of turning points (see Wasow [19]), the system

$$\varepsilon \frac{dc}{dt} = G(t)c \tag{1.24}$$

has no turning points in the interval $0 \leqq t \leqq 1$.

Then (see Turrittin [16]), the system (1.24) will have m linearly independent asymptotic solutions of the form

$$c_j(t, \varepsilon) = D_j(t, \varepsilon)\, e^{\lambda_j(0)\kappa}, \qquad j = 1, 2, \cdots, m, \tag{1.25}$$

and m asymptotic solutions of the form

$$c_j(t,\varepsilon) = D_j(1-t,\varepsilon)\,e^{\lambda_j(1)\sigma}, \qquad j = m+1, m+2, \cdots, 2m. \tag{1.26}$$

(Assumptions less restrictive than (H3) also admit these asymptotic solutions.)

(H4) With

$$D_j(0,0) = \begin{pmatrix} D_j^1 \\ D_j^2 \end{pmatrix}, \qquad j = 1, \cdots, 2m,$$

where the D_j^i's are m-dimensional vectors, the $m \times m$ matrices

$$D^1 = (D_1^1 D_2^1 \cdots D_m^1) \tag{1.27}$$

and

$$D^2 = ((D_{m+1}^2 - \pi_{30} D_{m+1}^1 \cdots (D_{2m}^2 - \pi_{30} D_{2m}^1)) \tag{1.28}$$

are nonsingular.

Once the asymptotic expansions of (1.12) are obtained, note that the control law (1.10) shows that we can determine the optimum control $u(t,\varepsilon)$ in the form

$$u(t,\varepsilon) = U(t,\varepsilon) + v(\kappa,\varepsilon) + w(\sigma,\varepsilon), \tag{1.29}$$

where the terms of v tend to zero as $\kappa \to \infty$ and the terms of w tend to zero as $\sigma \to \infty$. Further, the optimum cost $J^*(\varepsilon)$ will have an asymptotic power series expansion with J_0^* determined from the leading terms X_0, Z_0 and U_0 of the outer solution, i.e., J_0^* is the cost determined by the solution of the reduced problem.

We shall obtain the following statement of asymptotic correctness.

THEOREM. *Under the hypotheses* (H1)–(H4), *the optimal control problem* (1.1)–(1.3) *has a unique asymptotic solution for* ε *sufficiently small such that for every integer* $N \geqq 0$, *the optimal control and the corresponding trajectory satisfy*

$$\begin{aligned} u(t,\varepsilon) &= \sum_{l=0}^{N} (U_l(t) + v_l(\kappa) + w_l(\sigma))\varepsilon^l + O(\varepsilon^{N+1}), \\ x(t,\varepsilon) &= X_0(t) + \sum_{l=1}^{N} (X_l(t) + m_{1,l-1}(\kappa) + n_{1,l-1}(\sigma))\varepsilon^l + O(\varepsilon^{N+1}), \\ z(t,\varepsilon) &= \sum_{l=0}^{N} (Z_l(t) + m_{2l}(\kappa) + n_{2l}(\sigma))\varepsilon^l + O(\varepsilon^{N+1}) \end{aligned} \tag{1.30}$$

as $\varepsilon \to 0$ *uniformly throughout* $0 \leqq t \leqq 1$. *Here, the terms which are functions of* $\kappa = t/\varepsilon$ (*or* $\sigma = (1-t)/\varepsilon$) *decay to zero as* κ (σ) *tends to infinity* (*i.e., away from* $t = 0$ (*or* 1)). *Further, the optimal cost satisfies*

$$J^*(\varepsilon) = \sum_{l=0}^{N} J_l^* \varepsilon^l + O(\varepsilon^{N+1}) \quad \text{as } \varepsilon \to 0. \tag{1.31}$$

2. A simple example. We shall now obtain the asymptotic solution of the system

$$\frac{dx}{dt} = x,$$

$$\varepsilon\frac{dz}{dt} = -x - z + u$$

for scalars x, z and u on the interval $0 \leqq t \leqq 1$ with $x(0) = x_0$ and $z(0) = z_0$ prescribed, and with quadratic cost

$$J(\varepsilon) = \frac{1}{2}(x^2(1) + \varepsilon z^2(1)) + \frac{1}{2}\int_0^1 (x^2(t) + u^2(t))\,dt$$

to be minimized.

The appropriate Hamiltonian is

$$H(x, z, u, p_1, p_2) = \tfrac{1}{2}(x^2 + u^2) + p_1 x + p_2(-x - z + u),$$

and the Euler–Lagrange equations for minimum cost are

$$\frac{dx}{dt} = x,$$

$$\frac{dp_1}{dt} = -x - p_1 + p_2,$$

$$\varepsilon\frac{dz}{dt} = -x - z + u,$$

$$\varepsilon\frac{dp_2}{dt} = p_2,$$

$$0 = u + p_2.$$

The general solution of this constant coefficient linear system is

$$x(t) = -2c_2(\varepsilon)(1 + \varepsilon)\,e^t,$$

$$p_1(t) = c_1(\varepsilon)\,e^{-t} + c_2(\varepsilon)(1 + \varepsilon)\,e^t + 2\varepsilon c_3(\varepsilon)\exp\left[-\frac{1}{\varepsilon}(1 - t)\right],$$

$$z(t) = 2c_2(\varepsilon)\,e^t - c_3(\varepsilon)(1 + \varepsilon)\exp\left[-\frac{1}{\varepsilon}(1 - t)\right] + c_4(\varepsilon)\exp[-t/\varepsilon],$$

$$p_2(t) = 2c_3(\varepsilon)(1 + \varepsilon)\exp\left[-\frac{1}{\varepsilon}(1 - t)\right]$$

for arbitrary constants $c_i(\varepsilon)$. Applying the boundary conditions, neglecting only

asymptotically exponentially small terms, we have

$$c_1(\varepsilon) = \frac{3}{2}x_0\, e^2\left[1 + \frac{4\varepsilon}{9(1+\varepsilon)^2}\right],$$

$$c_2(\varepsilon) = \frac{-x_0}{2(1+\varepsilon)},$$

$$c_3(\varepsilon) = \frac{-x_0\, e}{3(1+\varepsilon)^2},$$

$$c_4(\varepsilon) = Z_0 + \frac{x_0}{1+\varepsilon}.$$

Thus, up to asymptotically exponentially small terms, we have

$$x(t) = x_0\, e^t,$$

$$z(t) = \frac{x_0}{1+\varepsilon}\left(-e^t + \frac{e}{3}\exp\left[-\frac{1}{\varepsilon}(1-t)\right] + e^{-t/\varepsilon}\right) + z_0\, e^{-t/\varepsilon},$$

$$u(t) = \frac{2}{3}\left(\frac{x_0\, e}{1+\varepsilon}\right)\exp\left[-\frac{1}{\varepsilon}(1-t)\right].$$

The minimum cost is then asymptotically given by

$$J^*(\varepsilon) = \frac{x_0^2}{2}\left[e^2\left(1 + \frac{4\varepsilon}{9(1+\varepsilon)^2}\right) + \int_0^1 e^{2t}\,dt + \frac{4e^2}{9}\,\frac{\varepsilon}{(1+\varepsilon)^2}\int_0^\infty e^{-2\sigma}\,d\sigma\right]$$

$$= \frac{x_0^2}{4}\left(3e^2 - 1 + \frac{4e^2}{9}\,\frac{\varepsilon}{(1+\varepsilon)^2}\right).$$

Note that here the reduced problem has the solution

$$X_0(t) = x_0\, e^t,$$

$$Z_0(t) = -x_0\, e^t,$$

$$U_0(t) = 0$$

and the corresponding cost

$$J^*(0) = \frac{x_0^2}{4}(3e^2 - 1).$$

3. Construction of the asymptotic expansions. The first term of the outer expansion (1.13) is given by the unique solution of the reduced problem (1.22). Higher order terms will be obtained by substituting the expansions for (1.13) into the system (1.11) and equating coefficients of terms of like order. Thus, they will satisfy nonhomogeneous forms of (1.18). By hypothesis (H1), G(t) is nonsingular, so the Z_j's and P_{2j}'s can be solved for successively as linear functions of X_j and P_{1j}. X_j and P_{1j} will then satisfy nonhomogeneous forms of the linear system (1.20). Further, $X_j(0)$ and $P_{1j}(1)$ will be known successively from (1.7), (1.9) and (1.12) in terms of the $X_k(0)$'s and $m_{1k}(0)$'s and the $P_{1k}(1)$'s and $\gamma_{1k}(0)$'s for $k < j$,

respectively. Since X_0 and P_{10} were uniquely determined by their boundary conditions, it follows that each $X_j(t)$ and $P_{1j}(t)$ will also be uniquely determined. Thus, the terms of the outer expansion can be determined successively once lower order terms in the expansions $m_1(0, \varepsilon)$ and $\gamma_1(0, \varepsilon)$ are known.

Since the outer expansion (1.13) satisfies the system (1.11) and since the boundary layer correction at $t = 1$ is asymptotically negligible near $t = 0$, (1.11) and (1.12) imply that the boundary layer correction at $t = 0$ satisfies the linear system

$$\begin{aligned}
\frac{dm_1}{d\kappa} &= \varepsilon A_1(\varepsilon\kappa, \varepsilon)m_1 + A_2(\varepsilon\kappa, \varepsilon)m_2 - \varepsilon S_1(\varepsilon\kappa, \varepsilon)\rho_1 - S(\varepsilon\kappa, \varepsilon)\rho_2, \\
\frac{d\rho_1}{d\kappa} &= -\varepsilon Q_1(\varepsilon\kappa, \varepsilon)m_1 - Q_2(\varepsilon\kappa, \varepsilon)m_2 - \varepsilon A_1'(\varepsilon\kappa, \varepsilon)\rho_1 - A_3'(\varepsilon\kappa, \varepsilon)\rho_2, \\
\frac{dm_2}{d\kappa} &= \varepsilon A_3(\varepsilon\kappa, \varepsilon)m_1 + A_4(\varepsilon\kappa, \varepsilon)m_2 - \varepsilon S'(\varepsilon\kappa, \varepsilon)\rho_1 - S_2(\varepsilon\kappa, \varepsilon)\rho_2, \\
\frac{d\rho_2}{d\kappa} &= -\varepsilon Q_2(\varepsilon\kappa, \varepsilon)m_1 - Q_3(\varepsilon\kappa, \varepsilon)m_2 - \varepsilon A_2'(\varepsilon\kappa, \varepsilon)\rho_1 - A_4'(\varepsilon\kappa, \varepsilon)\rho_2.
\end{aligned} \tag{3.1}$$

Substituting asymptotic power series for m_1, m_2, ρ_1 and ρ_2 into (3.1) and equating coefficients, we obtain a system of differential equations for the terms of this boundary layer correction. In particular, lowest order terms will satisfy the linear system

$$\begin{aligned}
\frac{dm_{10}}{d\kappa} &= A_{20}(0)m_{20} - S_0(0)\rho_{20}, \\
\frac{d\rho_{10}}{d\kappa} &= -Q_{20}(0)m_{20} - A_{30}'(0)\rho_{20}, \\
\frac{dm_{20}}{d\kappa} &= A_{40}(0)m_{20} - S_{20}(0)\rho_{20}, \\
\frac{d\rho_{20}}{d\kappa} &= -Q_{30}(0)m_{20} - A_{40}'(0)\rho_{20}.
\end{aligned} \tag{3.2}$$

Thus, m_{20} and ρ_{20} satisfy a linear system of order $2m$. By hypothesis (H2), we have

$$\begin{pmatrix} m_{20}(\kappa) \\ \rho_{20}(\kappa) \end{pmatrix} = \sum_{i=1}^{2m} \beta_i^0 \tilde{D}_i e^{\lambda_i(0)\kappa} \tag{3.3}$$

for some constants β_i^0. In order to eliminate exponentially growing solutions, hypothesis (H2) requires us to take

$$\beta_{m+1}^0 = \beta_{m+2}^0 = \cdots = \beta_{2m}^0 = 0.$$

Further, since the m-vector $m_{20}(0)$ is prescribed as

$$m_{20}(0) = z^0(0) - Z_0(0), \tag{3.4}$$

we have m equations for the m unknowns $\beta_1^0, \beta_2^0, \cdots, \beta_m^0$. The solution will be uniquely determined by (3.4), however, since $\tilde{D}_i = D_i(0,0)$, $i = 1, \cdots, m$, implies that the matrix of coefficients is D^1 which is nonsingular by hypothesis (H4). Once the β_i^0's are specified, note that (3.2) and (3.3) determine $\rho_{20}(\kappa)$, $dm_{10}/d\kappa$ and $d\rho_{10}/d\kappa$ completely. Because m_{10} and ρ_{10} tend to zero as $\kappa \to \infty$, we also have

$$
\begin{aligned}
m_{10}(\kappa) &= -\int_\kappa^\infty \frac{dm_{10}(s)}{d\kappa}\, ds, \\
\rho_{10}(\kappa) &= -\int_\kappa^\infty \frac{d\rho_{10}(s)}{d\kappa}\, ds.
\end{aligned} \tag{3.5}
$$

Thus, the terms $m_{10}, \rho_{10}, m_{20}$ and ρ_{20} all decay exponentially as $\kappa \to \infty$.

The system of differential equations satisfied by higher order terms of this boundary layer correction is obtained by successively equating like coefficients in (3.1). Thus,

$$
\begin{aligned}
\frac{dm_{2j}}{d\kappa} &= A_{40}(0)m_{2j} - S_{20}(0)\rho_{2j} + \mathscr{M}_{j-1}(\kappa), \\
\frac{d\rho_{2j}}{d\kappa} &= -Q_{30}(0)m_{2j} - A'_{40}(0)\rho_{2j} + \mathscr{P}_{j-1}(\kappa),
\end{aligned} \tag{3.6}
$$

where the exponentially decaying terms $\mathscr{M}_{j-1}$ and $\mathscr{P}_{j-1}$ are known successively. The general solution of (3.6) is

$$
\begin{pmatrix} m_{2j}(\kappa) \\ \rho_{2j}(\kappa) \end{pmatrix} = \sum_{i=1}^{2m} \beta_i^j \tilde{D}_i\, e^{\lambda_i(0)\kappa} + \begin{pmatrix} M_{2j}(\kappa) \\ P_{2j}(\kappa) \end{pmatrix}, \tag{3.7}
$$

where the β_i^j's are arbitrary constants and

$$
\begin{pmatrix} M_{2j}(\kappa) \\ P_{2j}(\kappa) \end{pmatrix} = \sum_{i=1}^{m} \tilde{D}_i\, e^{\lambda_i(0)\kappa} \int_0^\kappa \frac{\Delta_i(s)}{\Delta(s)}\, ds + \sum_{i=m+1}^{2m} \tilde{D}_i\, e^{\lambda_i(0)\kappa} \int_\infty^\kappa \frac{\Delta_i(s)}{\Delta(s)}\, ds,
$$

where

$$
\begin{aligned}
\Delta(s) &= \det(\tilde{D}_1\, e^{\lambda_1(0)s} \cdots \tilde{D}_{2m}\, e^{\lambda_{2m}(0)s}) \\
&= \exp\left(\sum_{j=1}^{2m} \lambda_j(0)s\right)
\end{aligned}
$$

and $\Delta_i(s)$ is the determinant obtained from $\Delta(s)$ by replacing its jth column by the exponentially decaying vector $\begin{pmatrix} \mathscr{M}_{i-1}(s) \\ \mathscr{P}_{i-1}(s) \end{pmatrix}$. Thus,

$$
\frac{\Delta_i(s)}{\Delta(s)} = e^{-\lambda_i(0)s}\delta_i(s),
$$

where $\delta_i(s)$ is decaying. Hypothesis (H2), then, implies that $\begin{pmatrix} M_{2j}(\kappa) \\ P_{2j}(\kappa) \end{pmatrix}$ also decays exponentially as $\kappa \to \infty$. It also implies that we must pick $\beta_{m+1}^j = \beta_{m+2}^j = \cdots = \beta_{2m}^j = 0$ in order to avoid growing solutions (3.7). Note that the jth term of the

outer expansion determines the initial value

$$m_{2j}(0) = z_j^0(0) - Z_j(0), \tag{3.8}$$

and this m-vector, by hypothesis (H4), uniquely determines the m constants $\beta_1^j, \cdots, \beta_m^j$ and the terms $m_{2j}(\kappa)$, $\rho_{2j}(\kappa)$, $dm_{1j}/d\kappa$ and $d\rho_{1j}/d\kappa$. Setting

$$\begin{aligned} m_{1j}(\kappa) &= -\int_\kappa^\infty \frac{dm_{1j}(s)}{d\kappa}\, ds, \\ \rho_{1j}(\kappa) &= -\int_\kappa^\infty \frac{d\rho_{1j}(s)}{d\kappa}\, ds, \end{aligned} \tag{3.9}$$

we have uniquely determined the jth terms of the boundary layer correction at $t = 0$ as exponentially decaying terms.

The boundary layer correction at $t = 1$ is determined in analogous fashion. Note that (1.11) and (1.12) imply that

$$\begin{aligned} \frac{dn_1}{d\sigma} &= -\varepsilon A_1(1 - \varepsilon\sigma, \varepsilon)n_1 - A_2(1 - \varepsilon\sigma, \varepsilon)n_2 + \varepsilon S_1(1 - \varepsilon\sigma, \varepsilon)\gamma_1 + S(1 - \varepsilon\sigma, \varepsilon)\gamma_2, \\ \frac{d\gamma_1}{d\sigma} &= \varepsilon Q_1(1 - \varepsilon\sigma, \varepsilon)n_1 + Q_2(1 - \varepsilon\sigma, \varepsilon)n_2 + \varepsilon A_1'(1 - \varepsilon\sigma, \varepsilon)\gamma_1 + A_3'(1 - \varepsilon\sigma, \varepsilon)\gamma_2, \\ \frac{dn_2}{d\sigma} &= -\varepsilon A_3(1 - \varepsilon\sigma, \varepsilon)n_1 - A_4(1 - \varepsilon\sigma, \varepsilon)n_2 + \varepsilon S'(1 - \varepsilon\sigma, \varepsilon)\gamma_1 + S_2(1 - \varepsilon\sigma, \varepsilon)\gamma_2, \\ \frac{d\gamma_2}{d\sigma} &= \varepsilon Q_2'(1 - \varepsilon\sigma, \varepsilon)n_1 + Q_3(1 - \varepsilon\sigma, \varepsilon)n_2 + \varepsilon A_2'(1 - \varepsilon\sigma, \varepsilon)\gamma_1 + A_4'(1 - \varepsilon\sigma, \varepsilon)\gamma_2. \end{aligned} \tag{3.10}$$

Equating coefficients successively, we find that each pair n_{2j}, γ_{2j} must be of the form

$$\begin{pmatrix} n_{2j}(\sigma) \\ \gamma_{2j}(\sigma) \end{pmatrix} = \sum_{i=1}^{2m} \alpha_i^j \tilde{\tilde{D}}_i\, e^{-\lambda_i(1)\sigma} + \begin{pmatrix} N_{2j}(\sigma) \\ G_{2j}(\sigma) \end{pmatrix}, \tag{3.11}$$

where the α_i^j's are constants and

$$\begin{pmatrix} N_{2j}(\sigma) \\ G_{2j}(\sigma) \end{pmatrix} = \sum_{i=1}^{m} \left(\tilde{\tilde{D}}_i\, e^{-\lambda_i(1)\sigma} \int_\infty^\sigma \frac{\tilde{\Delta}_i(s)}{\tilde{\Delta}(s)}\, ds \right) + \sum_{i=m+1}^{2m} \left(\tilde{\tilde{D}}_i\, e^{-\lambda_i(1)\sigma} \int_0^\sigma \frac{\tilde{\Delta}_i(s)}{\tilde{\Delta}(s)}\, ds \right),$$

$$\tilde{\Delta}(s) = \det\left(\tilde{\tilde{D}}_1\, e^{-\lambda_1(1)s} \cdots \tilde{\tilde{D}}_{2m}\, e^{-\lambda_{2m}(1)s}\right)$$

and $\tilde{\Delta}_i(s)$ is obtained from $\tilde{\Delta}(s)$ by replacing its ith column by a successively known exponentially decaying vector $\begin{pmatrix} \mathcal{n}_{i-1}(s) \\ \mathcal{g}_{i-1}(s) \end{pmatrix}$ (zero, when $j = 0$). In order to prevent growing solutions, it is necessary to select $\alpha_1^j = \alpha_2^j = \cdots = \alpha_m^j = 0$. Further, the jth term of the outer expansion determines the m-vector

$$\gamma_{2j}(0) - \pi_{30} n_{2j}(0). \tag{3.12}$$

Thus, since hypothesis (H4) implies that the matrix D^2 is nonsingular, we can uniquely determine α^j_{m+1}, $\alpha^j_{m+2}, \cdots, \alpha^j_{2m}$. Knowing n_{2j} and γ_{2j} determines $dn_{1j}/d\sigma$ and $d\gamma_{1j}/d\sigma$ in terms of lower order coefficients. Thus, by setting

$$
\begin{aligned}
n_1(\sigma) &= -\int_\sigma^\infty \frac{dn_{1j}(s)}{d\sigma}\,ds, \\
\gamma_{1j}(\sigma) &= -\int_\sigma^\infty \frac{d\gamma_{1j}(s)}{d\sigma}\,ds
\end{aligned}
\tag{3.13}
$$

we find that the jth terms of the boundary layer correction at $t = 1$ are completely determined as exponentially decaying terms.

From (1.10) and the expansions (1.12) for the costate vectors p_1 and p_2, we have

$$
u(t,\varepsilon) = U(t,\varepsilon) + v(\kappa,\varepsilon) + w(\sigma,\varepsilon), \tag{3.14}
$$

where

$$
\begin{aligned}
U(t,\varepsilon) &= -R^{-1}(t,\varepsilon)(B_1'(t,\varepsilon)P_1(t,\varepsilon) + B_2'(t,\varepsilon)P_2(t,\varepsilon)), \\
v(\kappa,\varepsilon) &= -R^{-1}(\varepsilon\kappa,\varepsilon)(\varepsilon B_1'(\varepsilon\kappa,\varepsilon)\rho_1(\kappa,\varepsilon) + B_2'(\varepsilon\kappa,\varepsilon)\rho_2(\kappa,\varepsilon)), \\
w(\sigma,\varepsilon) &= -R^{-1}(1-\varepsilon\sigma,\varepsilon)(\varepsilon B_1'(1-\varepsilon\sigma,\varepsilon)\gamma_1(\sigma,\varepsilon) + B_2'(1-\varepsilon\sigma,\varepsilon)\gamma_2(\sigma,\varepsilon)).
\end{aligned}
$$

Then, using (1.3) and the expansions (1.12) for the state vectors x and z, we obtain an asymptotic expansion for the optimal cost

$$
J^*(\varepsilon) = \frac{1}{2}\lambda(\varepsilon) + \frac{1}{2}\int_0^1 L_1(t,\varepsilon)\,dt + \frac{\varepsilon}{2}\int_0^\infty L_2(\kappa,\varepsilon)\,d\kappa + \frac{\varepsilon}{2}\int_0^\infty L_3(\sigma,\varepsilon)\,d\sigma, \tag{3.15}
$$

where

$$
\begin{aligned}
\lambda(\varepsilon) &= (X'(1,\varepsilon) + \varepsilon m_1'(0,\varepsilon))(\pi_1(\varepsilon)(X(1,\varepsilon) + \varepsilon m_1(0,\varepsilon))) \\
&\quad + 2\varepsilon\pi_2(\varepsilon)(Z(1,\varepsilon) + m_2(0,\varepsilon)) \\
&\quad + \varepsilon(Z'(1,\varepsilon) + m_2'(0,\varepsilon))\pi_3(\varepsilon)(Z(1,\varepsilon) + m_2(0,\varepsilon)), \\
L_1(t,\varepsilon) &= X'(t,\varepsilon)Q_1(t,\varepsilon)X(t,\varepsilon) + 2X'(t,\varepsilon)Q_2(t,\varepsilon)Z(t,\varepsilon) \\
&\quad + Z'(t,\varepsilon)Q_3(t,\varepsilon)Z(t,\varepsilon) + U'(t,\varepsilon)R(t,\varepsilon)U(t,\varepsilon), \\
L_2(\kappa,\varepsilon) &= 2\varepsilon X'(\varepsilon\kappa,\varepsilon)Q_1(\varepsilon\kappa,\varepsilon)m_1(\kappa,\varepsilon) \\
&\quad + \varepsilon^2 m_1'(\kappa,\varepsilon)Q_1(\varepsilon\kappa,\varepsilon)m_1(\kappa,\varepsilon) + 2X'(\varepsilon\kappa,\varepsilon)Q_2(\varepsilon\kappa,\varepsilon)m_2(\kappa,\varepsilon) \\
&\quad + 2\varepsilon m_1'(\kappa,\varepsilon)Q_2(\varepsilon\kappa,\varepsilon)Z(\varepsilon\kappa,\varepsilon) + 2\varepsilon m_1'(\kappa,\varepsilon)Q_2(\varepsilon\kappa,\varepsilon)m_2(\kappa,\varepsilon) \\
&\quad + 2Z'(\varepsilon\kappa,\varepsilon)Q_3(\varepsilon\kappa,\varepsilon)m_2(\kappa,\varepsilon) + m_2'(\kappa,\varepsilon)Q_3(\varepsilon\kappa,\varepsilon)m_2(\kappa,\varepsilon) \\
&\quad + 2U'(\varepsilon\kappa,\varepsilon)R(\varepsilon\kappa,\varepsilon)v(\kappa,\varepsilon) + v'(\kappa,\varepsilon)R(\varepsilon\kappa,\varepsilon)v(\kappa,\varepsilon)
\end{aligned}
$$

and

$$\begin{aligned} L_3(\sigma,\varepsilon) = {} & 2\varepsilon X'(1-\varepsilon\sigma,\varepsilon)Q_1(1-\varepsilon\sigma,\varepsilon)n_1(\sigma,\varepsilon) \\ & + \varepsilon^2 n_1'(\sigma,\varepsilon)Q_1(1-\varepsilon\sigma,\varepsilon)n_1(\sigma,\varepsilon) + 2X'(1-\varepsilon\sigma,\varepsilon)Q_2(1-\varepsilon\sigma,\varepsilon)n_2(\sigma,\varepsilon) \\ & + 2\varepsilon n_1'(\sigma,\varepsilon)Q_2(1-\varepsilon\sigma,\varepsilon)Z(1-\varepsilon\sigma,\varepsilon) + 2\varepsilon n_1'(\sigma,\varepsilon)Q_2(1-\varepsilon\sigma,\varepsilon)n_2(\sigma,\varepsilon) \\ & + 2Z'(1-\varepsilon\sigma,\varepsilon)Q_3(1-\varepsilon\sigma,\varepsilon)n_2(\sigma,\varepsilon) + n_2'(\sigma,\varepsilon)Q_3(1-\varepsilon\sigma,\varepsilon)n_2(\sigma,\varepsilon) \\ & + 2U'(1-\varepsilon\sigma,\varepsilon)R(1-\varepsilon\sigma,\varepsilon)w(\sigma,\varepsilon) + w'(\sigma,\varepsilon)R(1-\varepsilon\sigma,\varepsilon)w(\sigma,\varepsilon). \end{aligned}$$

Note that in (3.15), we have used the fact that products of exponentially decaying terms in κ and exponentially decaying terms in σ are negligible throughout $0 \leqq t \leqq 1$. Note, too, that because of the form of π_0, the boundary layer terms depending on κ and σ have no influence on the lowest order term J_0^* in the asymptotic expansion for the optimal cost $J^*(\varepsilon)$. Although boundary layer terms do not affect the zero order term of the state vector x, we note that they do give zero order contributions to the state vector z and the control vector u. These are significant only in the "boundary layer regions" near $t = 0$ and $t = 1$, however. Elsewhere, the outer solution is asymptotically valid.

In control theory, the outer solution corresponds to a "low order design" (cf. Kokotović and Sannuti [7]) and the singular perturbation technique which we have developed allows an improvement in the results of this low-order design.

We note that an alternate way of constructing the asymptotic solution is by use of the Riccati gain matrix (cf. Sannuti and Kokotović [14]). We have avoided this since our method is more direct and eliminates consideration of nonlinear equations. Our formulation is, then, an "open-loop" one as opposed to the "feedback" approach.

4. Proof of asymptotic correctness. A necessary and sufficient condition for optimality is that the Euler–Lagrange equations (1.11) be satisfied throughout $0 \leqq t \leqq 1$ together with the boundary conditions (1.7) and (1.9). In the absence of turning points (hypothesis (H3)), Turrittin [16] showed that the linear system (1.11) has $2n$ linearly independent asymptotic solutions of the form

$$F_j(t,\varepsilon), \qquad j = 1, 2, \cdots, 2n, \tag{4.1}$$

where F_{j0} is a solution of the reduced system (1.22); it has (under hypothesis (H2)) m asymptotic solutions of the form

$$F_j(t,\varepsilon)\exp\left[\frac{1}{\varepsilon}\int_0^t \lambda_{j-2n}(s)\,ds\right], \qquad j = 2n+1, \cdots, 2n+m, \tag{4.2}$$

where λ_{j-2n} is an eigenvalue of $G(t)$ with $\operatorname{Re}\lambda_{j-2n}(t) < 0$ throughout $0 \leqq t \leqq 1$; and it has m solutions of the form

$$F_j(t,\varepsilon)\exp\left[-\frac{1}{\varepsilon}\int_t^1 \lambda_{j-2n}(s)\,ds\right], \qquad j = 2n+m+1, \cdots, 2n+2m, \tag{4.3}$$

where $\lambda_{j-2n}(t)$ is an eigenvalue of $G(t)$ whose real part is negative in $[0,1]$. Note

that the solutions (4.2) and (4.3) can be rewritten in the forms

$$E_j(t, \varepsilon)\, e^{\lambda_{j-2n}(0)\kappa}, \qquad j = 2n + 1, \cdots, 2n + m, \tag{4.4}$$

and

$$E_j(t, \varepsilon)\, e^{-\lambda_{j-2n}(1)\sigma}, \qquad j = 2n + m + 1, \cdots, 2n + 2m, \tag{4.5}$$

which decay as either κ or σ tends to infinity. Thus, the general solution of the linear system (1.11) has the form

$$\begin{aligned} h(t, \varepsilon) = &\sum_{j=1}^{2n} c_j(\varepsilon) F_j(t, \varepsilon) \\ &+ \sum_{j=2n+1}^{2n+m} c_j(\varepsilon) E_j(\varepsilon\kappa, \varepsilon)\, e^{\lambda_{j-2n}(0)\kappa} \\ &+ \sum_{j=2n+m+1}^{2n+2m} c_j(\varepsilon) E_j(1 - \varepsilon\sigma, \varepsilon)\, e^{-\lambda_{j-2n}(1)\sigma}, \end{aligned} \tag{4.6}$$

where

$$h(t, \varepsilon) = \begin{pmatrix} x(t, \varepsilon) \\ p_1(t, \varepsilon) \\ z(t, \varepsilon) \\ p_2(t, \varepsilon) \end{pmatrix}$$

and the $c_j(\varepsilon)$'s are arbitrary coefficients. The $2n$ initial conditions (1.9) and the $2m$ terminal conditions (1.7) will provide $2n + 2m$ linear equations for the unknowns $c_1(\varepsilon), \cdots, c_{2n+2m}(\varepsilon)$. The determinant of coefficients will have an asymptotic expansion as $\varepsilon \to 0$ with nonzero leading term under the conditions (i) that the reduced problem has a unique solution (hypothesis (H1)) and (ii) that the matrices D^1 and D^2 are nonsingular (hypothesis (H4)) (cf. Wasow [18] or O'Malley and Keller [12] for an analogous detailed calculation). It follows that asymptotic expansions for the coefficients $c_j(\varepsilon)$ can be uniquely obtained (by Cramer's rule, for example). The resulting expansions will agree with (1.12) and the expansions for $u(t, \varepsilon)$ and $J^*(\varepsilon)$ follow directly from (1.5) and (1.3).

In practice, the differentiability assumptions and the assumptions of complete asymptotic expansions for the coefficients are too stringent. However, uniform asymptotic approximations of optimal solutions are needed, not complete expansions. Restricting attention to N-term approximations (as in the theorem), finite differentiability and asymptotic approximations of coefficients will clearly suffice.

5. Further problems. Three further problems are immediately obvious.

(a) *Several parameter problems.* In models for practical control problems, one must expect several small interrelated parameters (cf. Sannuti and Kokotović [15]). Singular perturbation techniques appropriate for such problems are developed in O'Malley [9], [11].

(b) *Time-invariant problems.* Linear regulator problems for time-invariant systems on infinite time intervals demand a separate investigation. Controllability assumptions are needed (cf. Kalman, Falb and Arbib [6]) and the singular pertur-

bation problem is more subtle (cf. Hoppensteadt [4]). Considerable progress has been made on this problem using the Riccati matrix formulation by Kokotović and Yackel [8].

(c) *Nonlinear problems.* The boundary layer method presented here can be extended to nonlinear two-point boundary value problems (cf. Hadlock [1], Hoppensteadt [5] and O'Malley [10], [11]). Results are not as explicit, however, and proofs can no longer rely on a constructed set of linearly independent solutions. Much progress on this problem has been made by Kokotović and his students (cf. Sannuti and Kokotović [15], Kokotović and Sannuti [7], Hadlock [1] and Sannuti [13]). Kokotović and Sannuti [15] give an interesting nonlinear model for a speed tracking system which again indicates the utility of asymptotic methods in practice.

REFERENCES

[1] C. R. HADLOCK, *Singular perturbations of a class of two-point boundary value problems arising in optimal control*, Rep. R-481, Coordinated Science Laboratory, University of Illinois, Urbana, 1970.

[2] C. HADLOCK, M. JAMSHIDI AND P. KOKOTOVIĆ, *Near optimum design of three time scale systems*, Proc. Fourth Annual Princeton Conference on Information Sciences and Systems, 1970, pp. 118–122.

[3] W. A. HARRIS, JR., *Singular perturbations of two point boundary problems*, J. Math. Mech., 11 (1962), pp. 371–382.

[4] F. C. HOPPENSTEADT, *Singular perturbations on the infinite interval*, Trans. Amer. Math. Soc., 123 (1966), pp. 521–535.

[5] ——— *Structure of decaying solutions of singular perturbation problems*, Proc. Eighth Annual Allerton Conf. on Circuit and Systems Theory, University of Illinois, Urbana, 1970, pp. 303–309.

[6] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[7] P. V. KOKOTOVIĆ AND P. SANNUTI, *Singular perturbation method for reducing the model order in optimal control design*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 377–384.

[8] P. V. KOKOTOVIĆ AND R. A. YACKEL, *Singular perturbation theory of linear state regulators*, Proc. Eighth Annual Allerton Conf. on Circuit and Systems Theory, 1970, University of Illinois, Urbana, pp. 310–321.

[9] R. E. O'MALLEY, JR., *Boundary value problems for linear systems of ordinary differential equations involving many small parameters*, J. Math. Mech., 18 (1969), pp. 835–856.

[10] ——— *Singular perturbation of a boundary value problem for a system of nonlinear differential equations*, J. Differential Equations, 8 (1970), pp. 431–447.

[11] ———, *Boundary layer methods for initial value problems*, SIAM Rev., 13 (1971), pp. 425–434.

[12] R. E. O'MALLEY, JR. AND J. B. KELLER, *Loss of boundary conditions in the asymptotic solution of linear ordinary differential equations. II: Boundary value problems*, Comm. Pure Appl. Math., 21 (1968), pp. 263–270.

[13] P. SANNUTI, *Singular perturbations of optimal control of a class of nonlinear systems*, Preprint, Rutgers University, New Brunswick, N.J., 1970.

[14] P. SANNUTI AND P. V. KOKOTOVIĆ, *Near-optimum design of linear systems by a singular perturbation method*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 15–22.

[15] ———, *Singular perturbation method for near optimum design of high-order non-linear systems*, Automatica, 5 (1969), pp. 773–779.

[16] H. L. TURRITTIN, *Asymptotic expansions of solutions of systems of ordinary differential equations containing a parameter*, Contrib. Theory Nonlinear Oscillations, II (1952), pp. 81–116.

[17] A. B. VASIL'EVA, *Asymptotic expansions of solutions of singularly perturbed problems*, Rep. 1077, Mathematics Research Center, University of Wisconsin, Madison, 1969.

[18] W. WASOW, *On the asymptotic solution of boundary value problems for ordinary differential equations containing a parameter*, J. Math. and Phys., 23 (1944), pp. 173–183.

[19] ———, *Asymptotic Expansions for Ordinary Differential Equations*, Interscience, New York, 1965.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

EFFICIENCY PRICES FOR OPTIMAL CONSUMPTION PLANS. IV*

BEZALEL PELEG†

Abstract. We consider a one-good economy which is described by a production function and a utility function. Cass and Yaari [1] distinguish several types of efficiency prices for efficient consumption plans in such a model. Motivated by their results, we characterize efficiency prices of the lim sup type (efficiency prices of type V). Our first result is that, under suitable assumptions on the production function, an intertemporal profit maximizing price system is of type V if and only if it keeps the present value of capital bounded throughout time. Our second result is, essentially, that a consumption plan is optimal if and only if it has a competitive price system which keeps the present value of capital bounded throughout time.

1. Introduction. We consider a one-good economy which is described by a production function $f(x)$ and a utility function $u(y)$. $f(x)$ is assumed to be a concave function of x for $x \geqq 0$ and to have some additional properties (see (i)–(iv) of § 2.) Some of our results (Theorems 3.11 and 4.16) require further assumptions on f. $u(y)$ is assumed to be a cancave and increasing function of y for $y \geqq 0$. A characterization of a certain system of efficiency prices for efficient consumption plans is given in § 3. Our investigation is motivated by the discussion in [1], and may be considered as a continuation of the investigation in [1]. A characterization of optimal consumption plans in terms of their systems of competitive prices is given in § 4. Our investigation of optimal consumption plans may be considered as a continuation of the investigation in [2]. In § 2, we collect some facts on concave functions. The reader is advised to skim through it and then use it as a reference for §§ 3 and 4. Our notation is adapted from that of [1]. Our terminology is borrowed from all our three references.

1.1. Why is the investigation of efficient and optimal plans interesting? First, since every inefficient program is inferior to some efficient program, the only programs that will be produced are the efficient ones. Hence, one looks for criteria of efficiency. Section 3 of this paper is devoted to an intrinsic characterization of a certain class of efficient programs, i.e., a characterization of such programs in terms of their capital sequences.

Secondly, when we add the utility function, then we have a maximization problem: A planner has to choose an optimal program, i.e., a program which maximizes utility, in an appropriate sense (see Definition 4.1), in the class of all feasible programs. This is a problem, of a particular type, of concave programming in an infinite-dimensional space. In § 4 of our paper, we give necessary and sufficient conditions for optimality in our model.

2. Some properties of concave functions. Let $f(x)$ be a concave function which is defined for $x \geqq 0$. The right-hand derivative of f, $f'_+(x)$, is defined for $x \geqq 0$ (the case $f'_+(0) = \infty$ is not excluded). The left-hand derivative of f, $f'_-(x)$, is defined for $x > 0$, and $f'_-(x) \geqq f'_+(x)$ for $x > 0$. Both f'_+ and f'_- are nonincreasing and if $x < y$, then $f'_+(x) \geqq f'_-(y)$.

* Received by the editors July 14, 1970, and in revised form April 11, 1971.

† Department of Mathematics, The Hebrew University of Jerusalem, Jerusalem, Israel.

DEFINITION 2.1. d is a *subderivative* of f at x if

$$f(y) - f(x) \leqq d(y - x) \qquad \text{for all } y \geqq 0.$$

If $x > 0$, then d is a subderivative of f at x if and only if $f'_-(x) \geqq d \geqq f'_+(x)$. d is a subderivative at 0 if and only if $\infty > d \geqq f'_+(0)$.

LEMMA 2.2. *Let $\underset{\sim}{x}$ be a sequence of real numbers which satisfies $x(t+1) < x(t)$, $t = 0, 1, 2, \cdots$, and $x(t) \geqq M > 0$, $t = 0, 1, 2, \cdots$, and let $x = \lim x(t)$. Let $d(t)$ be a subderivative of f at $x(t)$, $t = 0, 1, 2, \cdots$. Then $\lim d(t) = f'_+(x)$.*

Proof. $d(t) \leqq f'_-(x(t)) \leqq f'_+(x)$, $t = 0, 1, 2, \cdots$. Furthermore, $d(t+1) \geqq f'_+(x(t+1)) \geqq f'_-(x(t)) \geqq d(t)$, $t = 0, 1, 2, \cdots$. Hence, $d = \lim d(t)$ exists and $d \leqq f'_+(x)$. Let $y \geqq 0$. Then

$$f(y) - f(x(t)) \leqq d(t)(y - x(t)), \qquad t = 0, 1, 2, \cdots.$$

Hence, $f(y) - f(x) \leqq d(y - x)$. Thus, d is a subderivative of f at x and therefore $d \geqq f'_+(x)$. Hence, $d = f'_+(x)$.

LEMMA 2.3. *Let $\underset{\sim}{x}$ be a sequence of real numbers which satisfies $x(t+1) > x(t)$, $t = 0, 1, 2, \cdots$, and $0 \leqq x(t) \leqq M$, $t = 0, 1, 2, \cdots$, and let $x = \lim x(t)$. Let $d(t)$ be a subderivative of f at $x(t)$, $t = 0, 1, 2, \cdots$. Then $\lim d(t) = f'_-(x)$.*

The proof of Lemma 2.3 is similar to the proof of Lemma 2.2; hence, it will be omitted.

DEFINITION 2.4. f is *strictly concave at a point* x if

$$f(wx + (1 - w)y) > wf(x) + (1 - w)f(y)$$

$$\text{for all } y \geqq 0,\ y \neq x \quad \text{and} \quad 0 < w < 1.$$

f is *strictly concave* if it is strictly concave for all $x \geqq 0$.

LEMMA 2.5. *Let f be strictly concave at x. If d is a subderivative of f at x and $y \neq x$, then*

$$f(y) - f(x) < d(y - x).$$

Proof. Assume, on the contrary, that $f(y) = f(x) + d(y - x)$. Let $z = x/2 + y/2$. Then

$$\begin{aligned} f(x) + d(z - x) &\geqq f(z) > f(x)/2 + f(y)/2 \\ &= f(x)/2 + (f(x) + d(y - x))/2 = f(x) + d(z - x), \end{aligned}$$

which is impossible.

LEMMA 2.6. *The limit $\lim_{x \to 0+} f(x) = a$ exists and satisfies $a \geqq f(0)$. The function g which is defined by $g(x) = f(x)$ for $x > 0$ and $g(0) = a$ is continuous and concave. If f is strictly concave, then g is strictly concave.*

The proof, which is straightforward, is omitted.

LEMMA 2.7. *Let f be strictly concave and let $0 < M \leqq K$. Then there exists an $L < 0$ such that if $M \leqq x \leqq K$, $0 \leqq y \leqq x - M$, and d is a subderivative of f at x, then*

$$f(y) - f(x) - d(y - x) \leqq L.$$

Proof. Define g as in Lemma 2.6. g is continuous and strictly concave. We shall show that there exists an $L < 0$ such that if $M \leqq x \leqq K$, $0 \leqq y \leqq x - M$

and d is a subderivative of g at x, then

$$g(y) - g(x) - d(y - x) \leqq L.$$

Since $g(0) \geqq f(0)$ and $g(x) = f(x)$ for $x > 0$ this will prove the lemma. Assume now, on the contrary, that there exists no $L < 0$ with the above properties. Then there exist sequences $\underset{\sim}{x}$, $\underset{\sim}{y}$, and $\underset{\sim}{d}$ such that $M \leqq x(t) \leqq K$, $0 \leqq y(t) \leqq x(t) - M$, $d(t)$ is a subderivative of g at $x(t)$, $t = 0, 1, 2, \cdots$, and

$$\lim (g(y(t)) - g(x(t)) - d(t)(y(t) - x(t))) = 0.$$

The sequences $\underset{\sim}{x}$ and $\underset{\sim}{y}$ are bounded. Hence, we may assume that the limits $x = \lim x(t)$ and $y = \lim y(t)$ exist, $0 \leqq y \leqq x - M$. Let

$$L(t) = g(y(t)) - g(x(t)) - d(t)(y(t) - x(t)),$$

$$d(t) = (g(y(t)) - g(x(t)) - L(t))/(y(t) - x(t)).$$

Hence, $d = \lim d(t)$ exists and $g(y) - g(x) = d(y - x)$. Now, let $z \geqq 0$. Then

$$g(z) - g(x(t)) \leqq d(t)(z - x(t)), \qquad t = 0, 1, 2, \cdots.$$

Hence, $g(z) - g(x) \leqq d(z - x)$. Thus, d is a subderivative of g at x, and the equality $g(y) - g(x) = d(y - x)$ contradicts Lemma 2.5.

Henceforth, in this section, we shall assume:

(i) f is increasing, i.e., if $0 \leqq x < y$, then $f(y) > f(x)$.

LEMMA 2.8. *Every subderivative of f is positive and the limit* $\lim_{x \to \infty} f'_+(x) = f'_+(\infty)$ *exists.*

The proof, which is straightforward, is omitted.

We now add the following assumption:

(ii) $f'_+(\infty) < 1$.

LEMMA 2.9. *There exists a real number $x_1 \geqq 0$ such that $f(x) \leqq x$ for $x \geqq x_1$.*

The proof is omitted.

We now add two more assumptions:

(iii) $f(0) = 0$,

(iv) $f'_+(0) > 1$ and f is continuous at 0.

LEMMA 2.10. *The maximum*

$$y_* = \max \{f(x) - x | 0 \leqq x < \infty\} \tag{2.1}$$

exists and is positive.

The proof is omitted.

Define x_* by:

$$x_* = \min \{x | f(x) - x = y_*\}. \tag{2.2}$$

LEMMA 2.11. $x_* > 0$, $f'_-(x_*) \geqq 1 \geqq f'_+(x_*)$ *and* $f'_+(x) > 1$ *for* $0 \leqq x < x_*$.

The proof is omitted.

We shall use the following notation for sequences of real numbers. Let $\underset{\sim}{a}$ and $\underset{\sim}{b}$ be sequences of real numbers. We write $\underset{\sim}{a} \geqq \underset{\sim}{b}$ if $a(t) \geqq b(t)$, $t = 0, 1, 2, \cdots$. $\underset{\sim}{a} > \underset{\sim}{b}$ if $\underset{\sim}{a} \geqq \underset{\sim}{b}$ and $\underset{\sim}{a} \neq \underset{\sim}{b}$. $\underset{\sim}{a} \gg \underset{\sim}{b}$ if $a(t) > b(t)$, $t = 0, 1, 2, \cdots$. The zero sequence will be denoted by $\underset{\sim}{0}$. Thus, $\underset{\sim}{0} = (0, \cdots, 0, \cdots)$.

3. A characterization of a certain system of efficiency prices. Let $f(x)$ be a concave function which is defined for $x \geqq 0$ and which satisfies (i) and (iii) of § 2, i.e., f is increasing and $f(0) = 0$. We interpret f as the (stationary) production function of a one-good economy. Let $x_0 \geqq 0$ be the initial capital stock of the economy. $\underset{\sim}{x}$ is a *feasible capital sequence* if

$$x(0) = x_0 \quad \text{and} \quad 0 \leqq x(t+1) \leqq f(x(t)), \qquad t = 0, 1, 2, \cdots.$$

The set of all feasible capital sequences will be denoted by X. An $\underset{\sim}{x} \in X$ determines a *feasible consumption plan* $\underset{\sim}{y}$ by

$$y(t) = f(x(t)) - x(t+1), \qquad t = 0, 1, 2, \cdots.$$

We denote by Y the set of all feasible consumption plans. If $\underset{\sim}{x} \in X$ and $\underset{\sim}{y}$ is the feasible consumption plan which is determined by $\underset{\sim}{x}$, then the pair $(\underset{\sim}{x}, \underset{\sim}{y})$ will be called a *feasible program*. $\underset{\sim}{y} \in Y$ is *efficient* if there exists no $\underset{\sim}{y}^* \in Y$ such that $\underset{\sim}{y}^* > \underset{\sim}{y}$.

LEMMA 3.1. *Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program. If there exists a $t \geqq 0$ such that $x(t) = 0$, then $x(s) = 0$ for $s \geqq t$ and $\underset{\sim}{y}$ is efficient.*

Proof of Lemma 3.1. Assume, on the contrary, that there exists a feasible program $(\underset{\sim}{x}^*, \underset{\sim}{y}^*)$ such that $\underset{\sim}{y}^* > \underset{\sim}{y}$. Thus, $\underset{\sim}{y}^*(r) \geqq y(r)$, $r = 0, 1, 2, \cdots$, and there exists a natural number q, $0 \leqq q < \infty$, such that $y^*(q) > y(q)$. Since f is increasing,

$$(3.1) \qquad x^*(r) \leqq x(r) \quad \text{for } 0 \leqq r \leqq q \quad \text{and} \quad x^*(r) < x(r) \quad \text{for } r \geqq q+1.$$

Since $x(s) = 0$ for $s \geqq t$, the inequalities (3.1) contradict the feasibility of $\underset{\sim}{x}^*$.

Let $\underset{\sim}{x} \in X$. A sequence $\underset{\sim}{p} > \underset{\sim}{0}$ is an *intertemporal profit maximizing*[1] (i.p.m.) price system for $\underset{\sim}{x}$ if

$$p(t)f(x(t)) - p(t-1)x(t) \geqq p(t)f(x) - p(t-1)x$$
$$\text{for all } x \geqq 0, \qquad \text{and for } t = 1, 2, \cdots.$$

LEMMA 3.2. *Let $\underset{\sim}{x} \in X$ and let $\underset{\sim}{p}$ be an i.p.m. price system for $\underset{\sim}{x}$. Then*:

(i) *Either $\underset{\sim}{p} \gg \underset{\sim}{0}$ or there exists a $t \geqq 0$ such that $p(s) > 0$ for $0 \leqq s \leqq t$, $p(s) = 0$ for $s > t$, and $x(s) = 0$ for $s > t$.*

(ii) *If $p(t) > 0$, $t \geqq 1$, then $p(t-1)/p(t)$ is a subderivative of f at $x(t)$.*

The proof is omitted.

Following the terminology in [3, Definition 4.5], we make the following definition.

DEFINITION 3.3. Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program. An i.p.m. price system $\underset{\sim}{p}$ for $\underset{\sim}{x}$ is a *system of efficiency prices of type*[2] V for $\underset{\sim}{y}$ if

$$\limsup \sum_{s=0}^{t} p(s)(y^*(s) - y(s)) \leqq 0$$

for all $\underset{\sim}{y}^* \in Y$.

[1] $p(t-1)x$ is the cost of capital x invested for production in period t. $p(t)f(x)$ is the value of production during period t. Hence, the difference, $p(t)f(x) - p(t-1)x$ is the profit from production in period t. Thus, the profit from production, with respect to the price system p, is maximized in each period t by choosing $x = x(t)$.

[2] For a detailed classification of efficiency prices, the reader is referred to [3, § IV]. Let us only mention that even in our model there exist several types of efficiency prices as is shown in [1]. For the lack of a better terminology, we use that of [3].

LEMMA 3.4. *Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program. Let $\underset{\sim}{p}$ be an i.p.m. price system for $\underset{\sim}{x}$. If there exists a $t \geqq 0$ such that $p(t+1) = 0$, then $\underset{\sim}{p}$ is a system of efficiency prices of type* V *for $\underset{\sim}{y}$ and $\underset{\sim}{y}$ is efficient.*

Proof. By (i) of Lemma 3.2, $p(s) = 0$ for $s > t$ and $x(t+1) = 0$. By Lemma 3.1, $\underset{\sim}{y}$ is efficient. Now let $(\underset{\sim}{x}^*, \underset{\sim}{y}^*)$ be any feasible program. Then

$$\begin{aligned}
\limsup \sum_{r=0}^{s} p(r)(y^*(r) - y(r)) &= \sum_{r=0}^{t} p(r)(y^*(r) - y(r)) \\
&= \sum_{r=0}^{t} p(r)(f(x^*(r)) - x^*(r+1) - (f(x(r)) - x(r+1))) \\
&= \sum_{r=1}^{t} (p(r)f(x^*(r)) - p(r-1)x^*(r) - (p(r)f(x(r)) \\
&\qquad - p(r-1)x(r))) - p(t)x^*(t+1) \\
&\leqq -p(t)x^*(t+1) \leqq 0.
\end{aligned}$$

LEMMA 3.5. *Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program and let $\underset{\sim}{p}$ be an i.p.m. price system for $\underset{\sim}{x}$. If $\underset{\sim}{p} \gg \underset{\sim}{0}$ and $\underset{\sim}{p}$ is a system of efficiency prices of type* V *for $\underset{\sim}{y}$, then $\underset{\sim}{y}$ is efficient.*

The proof is omitted.

We now add assumption (ii) of § 2, i.e., $f'_+(\infty) < 1$.

LEMMA 3.6. *There exists a real number K such that if $\underset{\sim}{x} \in X$, then $x(t) \leqq K$ for $t = 0, 1, 2, \cdots$.*

Proof. By Lemma 2.9 there exists a real number x_1 such that $f(x) \leqq x$ for $x \geqq x_1$. Let $K = \max(x_0, x_1)$. Let $\underset{\sim}{x} \in X$. We shall prove by induction that $x(t) \leqq K$ for $t = 0, 1, 2, \cdots$. For $t = 0$, $x(0) = x_0 \leqq K$. Assume now that $x(t) \leqq K$, $t \geqq 0$. If $x(t) \leqq x_1$, then

$$x(t+1) \leqq f(x(t)) \leqq f(x_1) \leqq x_1 \leqq K.$$

If $x(t) > x_1$, then

$$x(t+1) \leqq f(x(t)) \leqq x(t) \leqq K.$$

LEMMA 3.7. *Assume that f is strictly concave. Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program and let $\underset{\sim}{p}$ be an i.p.m. price system for $\underset{\sim}{x}$. If $\underset{\sim}{p}$ is bounded, then $\underset{\sim}{p}$ is a system of efficiency prices of type* V *for $\underset{\sim}{y}$ and $\underset{\sim}{y}$ is efficient.*

Proof. If there exists a $t \geqq 0$ such that $p(t+1) = 0$, then the proof follows from Lemma 3.4. Hence, we may assume that $\underset{\sim}{p} \gg \underset{\sim}{0}$. By (ii) of Lemma 3.2, $d(t) = p(t-1)/p(t)$ is a subderivative of f at $x(t)$ for $t = 1, 2, \cdots$. By assumption, there exists a real number R such that $p(t) \leqq R$ for $t = 0, 1, 2, \cdots$. Let $(\underset{\sim}{x}^*, \underset{\sim}{y}^*)$ be any feasible program and let $t \geqq 0$. Then

$$\begin{aligned}
S(t) &= \sum_{s=0}^{t} p(s)(y^*(s) - y(s)) \\
&= \sum_{s=0}^{t} p(s)(f(x^*(s)) - x^*(s+1) - (f(x(s)) - x(s+1))) \\
&= \sum_{s=1}^{t} p(s)(f(x^*(s)) - f(x(s)) - d(s)(x^*(s) - x(s))) \\
&\qquad + p(t)(x(t+1) - x^*(t+1)) \\
&\leqq p(t)(x(t+1) - x^*(t+1)).
\end{aligned}$$

Let $e > 0$ be given. Denote $M = e/R$. We distinguish the following cases.

(a) There exists a natural number t_0 such that

$$x(t+1) - x^*(t+1) < M \qquad \text{for } t \geqq t_0.$$

In this case, $S(t) < e$ for $t \geqq t_0$.

(b) The set $A = \{t | x(t+1) - x^*(t+1) \geqq M\}$ is infinite.

Let $A = \{t_1, t_2, \cdots, t_j, \cdots\}$, where $t_j < t_{j+1}$, $j = 1, 2, \cdots$. We distinguish the following subcases.

(b.1) $\lim p(t_j) = 0$.

By Lemma 3.6 there exists a real number K such that $x(t) \leqq K$ for $t = 0, 1, 2, \cdots$. There exists a natural number k such that if $j \geqq k$, then $p(t_j)K < e$. Let $t \geqq t_k$. If $t \notin A$, then

$$S(t) \leqq p(t)(x(t+1) - x^*(t+1)) < e,$$

and if $t \in A$, then

$$S(t) \leqq p(t)x(t+1) \leqq p(t)K < e.$$

(b.2) There exist a real number $N > 0$ and an infinite subset B of A such that $p(t) \geqq N$ for $t \in B$.

Let $B = \{s_1, s_2, \cdots, s_j, \cdots\}$, where $s_j < s_{j+1}$, $j = 1, 2, \cdots$; $x(s_j + 1) \geqq x^*(s_j + 1) + M \geqq M$, $j = 1, 2, \cdots$. Hence, $p(s_j + 1) = p(s_j)/d(s_j + 1) \geqq N/f'_-(M) = N_1$. By Lemma 2.7, there exists an $L < 0$ such that

$$f(x^*(s_j+1)) - f(x(s_j+1)) - d(s_j+1)(x^*(s_j+1) - x(s_j+1)) \leqq L,$$
$$j = 1, 2, \cdots.$$

There exists a natural number k such that $kLN_1 + RK < e$. If $t \geqq s_k + 1$, then

$$\begin{aligned} S(t) &\leqq \sum_{j=1}^{k} p(s_j+1)(f(x^*(s_j+1)) - f(x(s_j+1)) \\ &\quad - d(s_j+1)(x^*(s_j+1) - x(s_j+1))) + p(t)x(t+1) \\ &\leqq kLN_1 + RK < e. \end{aligned}$$

Thus, $\underset{\sim}{p}$ is a system of efficiency prices of type V for $\underset{\sim}{y}$. By Lemma 3.5, $\underset{\sim}{y}$ is efficient.

Example 3.8. Choose $0 < w < 1$ such that $2 > 1/(1-w)^3$. Let $a = (1-w)^2$ and $b = 1/(1-w)^2$. Define a function f by $f(x) = 2x^{1/2}$ for $0 \leqq x \leqq b$, and $f(x) = f(b) + (1-w)(x-b)$ for $x > b$. Then f is concave and satisfies assumptions (i)–(iv) of § 2. Let $x_0 = a$ and let $0 < v < f(a) - b$. Define a feasible capital sequence $\underset{\sim}{x}$ by $x(t) = a$ for t even, and $x(t) = b + v$ for t odd. The feasible consumption plan $\underset{\sim}{y}$ which is determined by $\underset{\sim}{x}$ is given by $y(t) = f(a) - b - v$ for t even, and $y(t) = f(b) + (1-w)v - a$ for t odd. Let $p(t) = 1 - w$ for t even, and $p(t) = 1$ for t odd. Then $\underset{\sim}{p}$ is a bounded i.p.m. price system for $\underset{\sim}{x}$. Consider the feasible program $(\underset{\sim}{x}^*, \underset{\sim}{y}^*)$, where $x^*(t) = a$ for t even, and $x^*(t) = b$ for t odd. $y^*(t) - y(t) = v$ for t even and $y^*(t) - y(t) = -(1-w)v$ for t odd. Hence, for t even,

$$\sum_{s=0}^{t} p(s)(y^*(s) - y(s)) = (1-w)v.$$

Thus, $\underset{\sim}{p}$ is not a system of efficiency prices of type V for $\underset{\sim}{y}$.

Remark 3.9. We conclude from Example 3.8 that the assumption that f is strictly concave is essential for the validity of Lemma 3.7. Notice that in the above example x_*, the golden rule capital stock, is equal to 1 (see (2.2)), and f is strictly concave near x_*.

We now add assumption (iv) of § 2, i.e., that $f'_+(0) > 1$ and f is continuous at 0.

LEMMA 3.10. *Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program and let $\underset{\sim}{p}$ be an i.p.m. price system for $\underset{\sim}{x}$. Define a sequence $\underset{\sim}{q}$ by $q(t) = p(t)x(t+1)$, $t = 0, 1, 2, \cdots$. Then $\underset{\sim}{p}$ is bounded if and only if $\underset{\sim}{q}$ is bounded.*

Proof. By Lemma 3.6 there exists a K such that $x(t) \leqq K$, $t = 0, 1, 2, \cdots$. Hence, if $\underset{\sim}{p}$ is bounded, then $\underset{\sim}{q}$ is bounded. Assume now that $\underset{\sim}{q}$ is bounded. Suppose, on the contrary, that $\underset{\sim}{p}$ is not bounded. Then there exists an increasing sequence $\{t_1, t_2, \cdots, t_j, \cdots\}$ of natural numbers such that $p(t_{j+1}) > p(t_j)$, $j = 1, 2, \cdots$, and $\lim p(t_j) = \infty$. By (i) of Lemma 3.2, $\underset{\sim}{p} \gg \underset{\sim}{0}$. By (ii) of Lemma 3.2, $p(t-1)/p(t)$ is a subderivative of f at $x(t)$, $t = 1, 2, \cdots$. Hence, by Lemma 2.11, we may assume that $x(t_j) \geqq x_*$, $j = 1, 2, \cdots$. For $j \geqq 2$, $p(t_j - 1) \geqq f'_+(x(t_j))p(t_j) \geqq f'_+(K)p(t_j)$. Thus,

$$p(t_j - 1)x(t_j) \geqq f'_+(K)x_*p(t_j), \qquad j \geqq 2,$$

which contradicts our assumption that $\underset{\sim}{q}$ is bounded.

THEOREM 3.11. *Assume that f is strictly concave. Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program and let $\underset{\sim}{p}$ be an i.p.m. price system for $\underset{\sim}{x}$. If there exists a real number R such that $p(t)x(t+1) \leqq R$, $t = 0, 1, 2, \cdots$, then $\underset{\sim}{p}$ is a system of efficiency prices of type* V *for $\underset{\sim}{y}$, and $\underset{\sim}{y}$ is efficient.*

Proof. The result follows from Lemma 3.7 and Lemma 3.10.

Theorem 3.11 tells us that under suitable assumptions on the production function (strict concavity and assumptions (i)–(iv) of § 2), an i.p.m. price system which keeps the present value of capital bounded throughout time is a system of efficiency prices of type V and guarantees efficiency. We now turn to a proof of a converse result.

LEMMA 3.12. *Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program and let $\underset{\sim}{p}$ be an i.p.m. price system for $\underset{\sim}{x}$. If there exist an increasing sequence of natural numbers $A = \{t_1, t_2, \cdots, t_j, \cdots\}$ and a real number $M > 0$ such that*

(a) $\lim p(t_j) = \infty$, *and*

(b) $y(t_j + 1) = f(x(t_j + 1)) - x(t_j + 2) \geqq M$, $j = 1, 2, \cdots$,

then $\underset{\sim}{p}$ is not a system of efficiency prices of type V *for $\underset{\sim}{y}$.*

Proof. Let $B = \{t_1 + 1, t_2 + 1, \cdots, t_j + 1, \cdots\}$. We define a real function $\underset{\sim}{z}$ on B by

$$f(z(t_j + 1)) = x(t_j + 2), \qquad j = 1, 2, \cdots.$$

We claim that there exists a real number $N > 0$ such that

(c) $x(t_j + 1) - z(t_j + 1) \geqq N$, $j = 1, 2, \cdots$.

Assume, on the contrary, that there exists an increasing subsequence $\{s_1, s_2, \cdots, s_j, \cdots\}$ of A such that $\lim (x(s_j + 1) - z(s_j + 1)) = 0$. By Lemma 3.6,

$\underset{\sim}{x}$ is bounded. Hence, we may assume that the limit $\lim x(s_j + 1)$ exists. Thus,

$$\begin{aligned} 0 &= \lim (f(x(s_j + 1)) - f(z(s_j + 1))) \\ &= \lim (f(x(s_j + 1)) - x(s_j + 2)), \end{aligned}$$

which contradicts (b).

We shall now define inductively an increasing subsequence $\{r_1, r_2, \cdots, r_j, \cdots\}$ of A. Let

$$r_1 = \min \{t | t \in A \text{ and } p(t)(x(t + 1) - z(t + 1)) \geqq 1\}.$$

By (a) and (c), r_1 is well-defined. It follows from (a) and (i) of Lemma 3.2 that $\underset{\sim}{p} \gg \underset{\sim}{0}$. Let $d(t) = p(t - 1)/p(t)$, $t = 1, 2, \cdots$. Assume now that $r_1, \cdots, r_j$ have been chosen, $j \geqq 1$, such that $r_l < r_{l+1}$, $l = 1, \cdots, j - 1$, and

$$\sum_{h=1}^{l-1} p(r_h + 1)(f(z(r_h + 1)) - f(x(r_h + 1)) - d(r_h + 1)(z(r_h + 1) - x(r_h + 1))) + p(r_l)(x(r_l + 1) - z(r_l + 1)) \geqq 1, \qquad l = 1, \cdots, j.$$

Define r_{j+1} by

$$r_{j+1} = \min \{t | t \in A, t \geqq r_j + 1, \text{ and } t \text{ satisfies (d)}\},$$

where

$$\text{(d)} \quad \sum_{h=1}^{j} p(r_h + 1)(f(z(r_h + 1)) - f(x(r_h + 1)) - d(r_h + 1)(z(r_h + 1) - x(r_h + 1))) + p(t)(x(t + 1) - z(t + 1)) \geqq 1.$$

By (a) and (c), r_{j+1} is well-defined. Let $C = \{r_1 + 1, r_2 + 1, \cdots, r_j + 1, \cdots\}$. We now define a sequence $\underset{\sim}{x}^*$ by $x^*(t) = x(t)$ if $t \notin C$ and $x^*(t) = z(t)$ if $t \in C$. We claim that $\underset{\sim}{x}^*$ is a feasible capital sequence. To see this, observe that $x^*(0) = x(0) = x_0$, and that if $t \notin C$, then

$$f(x^*(t)) - x^*(t + 1) = f(x(t)) - x^*(t + 1) \geqq f(x(t)) - x(t + 1) \geqq 0.$$

If $t \in C$, then

$$\begin{aligned} f(x^*(t)) - x^*(t + 1) &= f(z(t)) - x^*(t + 1) = x(t + 1) - x^*(t + 1) \\ &\geqq x(t + 1) - x(t + 1) = 0. \end{aligned}$$

Let $\underset{\sim}{y}^*$ be the feasible consumption plan which is determined by $\underset{\sim}{x}^*$. For $j = 1, 2, \cdots$ we have

$$\begin{aligned} \sum_{t=0}^{r_j} p(t)(y^*(t) - y(t)) &= \sum_{t=1}^{r_j} p(t)(f(x^*(t)) - f(x(t)) \\ &\qquad - d(t)(x^*(t) - x(t))) + p(r_j)(x(r_j + 1) - x^*(r_j + 1)) \\ &= \sum_{h=1}^{j-1} p(r_h + 1)(f(z(r_h + 1)) - f(x(r_h + 1)) - d(r_h + 1)(z(r_h + 1) \\ &\qquad - x(r_h + 1))) + p(r_j)(x(r_j + 1) - z(r_j + 1)) \geqq 1. \end{aligned}$$

Hence, $\underset{\sim}{p}$ is not a system of efficiency prices of type V for $\underset{\sim}{y}$.

THEOREM 3.13. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a feasible program and let* $\underset{\sim}{p}$ *be an i.p.m. price system for* $\underset{\sim}{x}$. *If* $\underset{\sim}{p}$ *is a system of efficiency prices of type* V *for* $\underset{\sim}{y}$, *then there exists a real number* R *such that* $p(t)x(t+1) \leqq R$, $t = 0, 1, 2, \cdots$.

Proof. By Lemma 3.10 it is sufficient to show that $\underset{\sim}{p}$ is bounded. Assume, on the contrary, that $\limsup p(t) = \infty$. We distinguish the following cases:

(a) There exists a t_0 such that if $t \geqq t_0$, then $p(t+1) \geqq p(t)$.

In this case $\lim p(t) = \infty$. Furthermore, $\underset{\sim}{p} \gg \underset{\sim}{0}$ and $p(t-1)/p(t)$ is a subderivative of f at $x(t)$, $t \geqq 1$. Hence, by Lemma 2.11, $x(t) \geqq x_*$ for $t > t_0$. We claim that $\limsup y(t) > 0$. Assume, on the contrary, that $\lim y(t) = 0$. Define a feasible capital sequence $\underset{\sim}{x}^*$ by $x^*(t) = x(t)$, $t \leqq t_0$, and $x^*(t) = x_*$, $t > t_0$. Let $\underset{\sim}{y}^*$ be the feasible consumption plan which is determined by $\underset{\sim}{x}^*$. $\underset{\sim}{y}^*(t) = y_*$ for $t > t_0$ (see (2.1)). Hence

$$\lim \sum_{s=0}^{t} p(s)(y^*(s) - y(s)) = \infty,$$

which contradicts our assumption that $\underset{\sim}{p}$ is a system of efficiency prices of type V for $\underset{\sim}{y}$. Hence there exist an increasing sequence $\{t_1, t_2, \cdots, t_j, \cdots\}$ of natural numbers and a real number $M > 0$ such that

$$y(t_j + 1) \geqq M, \qquad j = 1, 2, \cdots.$$

Since $\lim p(t_j) = \infty$, by Lemma 3.12, $\underset{\sim}{p}$ is not a system of efficiency prices of type V for $\underset{\sim}{y}$, which contradicts our assumption.

(b) There exists an increasing sequence of natural numbers $A = \{t_1, t_2, \cdots, t_j, \cdots\}$ such that $p(t_j + 1) < p(t_j)$, $j = 1, 2, \cdots$.

We may assume A is chosen such that there exists an increasing sequence of natural numbers $\{s_1, s_2, \cdots, s_j, \cdots\}$ which satisfies

(b.1) $t_j + 2 \leqq s_j \leqq t_{j+1} + 1$, $j = 1, 2, \cdots$,

(b.2) $p(s_j) > p(t_j + 2)$, $j = 1, 2, \cdots$,

(b.3) $\lim p(s_j) = \infty$.

By (i) of Lemma 3.2, $\underset{\sim}{p} \gg \underset{\sim}{0}$, and by (ii) of Lemma 3.2, $p(t-1)/p(t)$ is a subderivative of f at $x(t)$ for $t \geqq 1$. We now define an increasing sequence of natural numbers $\{r_1, r_2, \cdots, r_j, \cdots\}$ by:

(b.4) $r_j = \max\{t | t_j + 2 \leqq t \leqq t_{j+1} + 1 \text{ and } p(t) \geqq p(s), t_j + 2 \leqq s \leqq t_{j+1} + 1\}$.

By (b.2), $s_j > t_j + 2$. Hence, by (b.1), $t_{j+1} \geqq t_j + 2$. By (b), $p(t_{j+1} + 1) < p(t_{j+1})$. Hence, it follows from (b.4) that $p(r_j) > p(r_j + 1)$. Using Lemma 2.11, we conclude that $x(r_j + 1) \leqq x_*$. It follows from (b.2) and (b.4) that $p(r_j) \geqq p(r_j - 1)$. Hence, by Lemma 2.11, $x(r_j) \geqq x_*$. Thus,

(b.5) $y(r_j) = f(x(r_j)) - x(r_j + 1) \geqq f(x_*) - x_* = y_*$, $j = 1, 2, \cdots$.

It follows from (b.3) and (b.4) that $\lim p(r_j) = \infty$. By Lemma 3.6 there exists a real number K such that $x(t) \leqq K$, $t = 0, 1, 2, \cdots$. Thus,

$$p(r_j - 1) \geqq f'_+(x(r_j))p(r_j) \geqq f'_+(K)p(r_j), \qquad j = 1, 2, \cdots.$$

Hence,

(b.6) $\lim p(r_j - 1) = \infty$.

It follows from (b.5), (b.6) and Lemma 3.12 that $\underset{\sim}{p}$ is not a system of efficiency prices of type V for $\underset{\sim}{y}$, which contradicts our assumption.

The following two lemmas will be used in the next section.

LEMMA 3.14. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a feasible program and let* $\underset{\sim}{p}$ *be an i.p.m. price system for* $\underset{\sim}{x}$. *Assume that* $\underset{\sim}{p}$ *is bounded,* $x(t) > x_*$, $t = 0, 1, 2, \cdots$, *and* $\lim x(t) = x_*$. *Then* $\underset{\sim}{p}$ *is a system of efficiency prices of type* V *for* $\underset{\sim}{y}$.

Proof. $\underset{\sim}{x} \gg \underset{\sim}{0}$. Hence, by (i) of Lemma 3.2, $\underset{\sim}{p} \gg \underset{\sim}{0}$. By (ii) of Lemma 3.2, $d(t) = p(t-1)/p(t)$ is a subderivative of f at $x(t)$ for $t = 1, 2, \cdots$. Hence, $d(t) \leqq f'_+(x_*) \leqq 1$, $t = 1, 2, \cdots$. Thus, $\underset{\sim}{p}$ is a nondecreasing sequence. By assumption, there exists a real number R such that $p(t) \leqq R$, $t = 0, 1, 2, \cdots$. Let $(\underset{\sim}{x}^*, \underset{\sim}{y}^*)$ be any feasible program and let $t \geqq 0$. Then

$$S(t) = \sum_{s=0}^{t} p(s)(y^*(s) - y(s))$$

$$= \sum_{s=1}^{t} p(s)(f(x^*(s)) - f(x(s)) - d(s)(x^*(s) - x(s)))$$

$$+ p(t)(x(t+1) - x^*(t+1)).$$

Let $e > 0$ be given. Denote $M = e/R$. We distinguish the following cases:

(a) There exists a natural number t_0 such that $x(t+1) - x^*(t+1) < M$ for $t \geqq t_0$.

In this case $S(t) < e$ for $t \geqq t_0$.

(b) The set $A = \{t | x(t+1) - x^*(t+1) \geqq M\}$ is infinite.

Let $A = \{t_1, t_2, \cdots, t_j, \cdots\}$, where $t_j < t_{j+1}$, $j = 1, 2, \cdots$. It follows from Lemma 2.11 and the continuity of f that there exists a real number $L < 0$ such that

$$f(x) - f(x_*) - f'_+(x_*)(x - x_*) \leqq L \qquad \text{for } 0 \leqq x \leqq x_* - M/2. \tag{3.2}$$

It follows from Lemma 2.2 that $\lim d(t) = f'_+(x_*)$. By Lemma 3.6, there exists a real number K such that $x(t) \leqq K$, $t = 0, 1, 2, \cdots$. There exists a natural number k such that for $j \geqq k$,

$$x(t_j + 1) < x_* + \frac{M}{2}, \qquad f'_+(x_*) - d(t_j + 1) < \frac{-Lp(0)}{4KR},$$

and

$$|f(x_*) - f'_+(x_*)x_* - (f(x(t_j+1)) - d(t_j+1)x(t_j+1))| < \frac{-Lp(0)}{4R}.$$

From (3.2) it follows that

$$f(x^*(t_j+1)) - f(x_*) - f'_+(x_*)(x^*(t_j+1) - x_*) \leqq L, \qquad j \geqq k.$$

Choose $l \geqq k$ such that $(l-k+1)Lp(0)/2 + RK < e$. If $t \geqq t_l + 1$, then

$$\begin{aligned}
S(t) &\leqq \sum_{j=k}^{l} p(t_j+1)(f(x^*(t_j+1)) - f(x(t_j+1)) \\
&\qquad - d(t_j+1)(x^*(t_j+1) - x(t_j+1))) + p(t)x(t+1) \\
&= \sum_{j=k}^{l} p(t_j+1)(f(x^*(t_j+1)) - f(x_*) - f'_+(x_*)(x^*(t_j+1) - x_*)) \\
&\qquad + \sum_{j=k}^{l} p(t_j+1)(f(x_*) - f'_+(x_*)x_* - (f(x(t_j+1)) - d(t_j+1)x(t_j+1))) \\
&\qquad + \sum_{j=k}^{l} p(t_j+1)x^*(t_j+1)(f'_+(x_*) - d(t_j+1)) + p(t)x(t+1) \\
&\leqq (l-k+1)p(0)L - (l-k+1)Lp(0)/4 - (l-k+1)Lp(0)/4 + KR \\
&= (l-k+1)p(0)L/2 + RK < e.
\end{aligned}$$

LEMMA 3.15. *Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program and let $\underset{\sim}{p}$ be an i.p.m. price system for $\underset{\sim}{x}$. Assume that $\lim x(t) = x_*$, and that there exists an r such that $p(t+1) \leqq p(t)$ for $t \geqq r$. Then $\underset{\sim}{p}$ is a system of efficiency prices of type* V *for $\underset{\sim}{y}$.*

Proof. $\underset{\sim}{x} \gg \underset{\sim}{0}$. Hence, $\underset{\sim}{p} \gg \underset{\sim}{0}$. $d(t) = p(t-1)/p(t)$ is a subderivative of f at $x(t)$, $t = 1, 2, \cdots$. By assumption the limit $N = \lim p(t)$ exists. Let $(\underset{\sim}{x}^*, \underset{\sim}{y}^*)$ be any feasible program and let $t \geqq 0$. Then

$$\begin{aligned}
S(t) &= \sum_{s=0}^{t} p(s)(y^*(s) - y(s)) \\
&= \sum_{s=1}^{t} p(s)(f(x^*(s)) - f(x(s)) - d(s)(x^*(s) - x(s))) \\
&\qquad + p(t)(x(t+1) - x^*(t+1)).
\end{aligned}$$

Let $e > 0$ be given. Denote $M = e/p(r)$. We distinguish the following cases:

(a) $N = 0$.

By Lemma 3.6 there exists a K such that $x(t) \leqq K$, $t = 0, 1, 2, \cdots$. There exists a natural number k such that if $t \geqq k$, then $p(t)K < e$. Thus, if $t \geqq k$, then

$$S(t) \leqq p(t)x(t+1) \leqq p(t)K < e.$$

(b) $N > 0$.

$d(t) \geqq 1$ for $t \geqq r$. Hence, in this case, $\lim d(t) = 1$. We distinguish the following subcases:

(b.1) There exists a natural number $t_0 \geqq r$ such that

$$x(t+1) - x^*(t+1) < M \qquad \text{for } t \geqq t_0.$$

In this case if $t \geqq t_0$, then

$$S(t) \leqq p(t)(x(t+1) - x^*(t+1)) < p(r)M = e.$$

(b.2) $A = \{t | t \geqq r \text{ and } x(t+1) - x^*(t+1) \geqq M\}$ is infinite.

Let $A = \{t_1, t_2, \cdots, t_j, \cdots\}$, where $t_j < t_{j+1}$, $j = 1, 2, \cdots$. There exists a real number $L < 0$ such that

$$f(x) - f(x_*) - (x - x_*) \leqq L \qquad \text{for } 0 \leqq x \leqq x_* - M/2.$$

There exists a natural number k such that for $j \geqq k$,

$$x(t_j + 1) < x_* + M/2$$

and

$$|f(x_*) - x_* - (f(x(t_j + 1)) - d(t_j + 1)x(t_j + 1))| \leqq -LN/2p(r).$$

For $j \geqq k$, $x^*(t_j + 1) \leqq x_* - M/2$. Hence,

$$f(x^*(t_j + 1)) - f(x_*) - (x^*(t_j + 1) - x_*) \leqq L, \qquad j \geqq k.$$

Choose $l \geqq k$ such that $(l - k + 1)LN/2 + p(r)K < e$. If $t \geqq t_l + 1$, then

$$\begin{aligned} S(t) &\leqq \sum_{j=k}^{l} p(t_j + 1)(f(x^*(t_j + 1)) - f(x(t_j + 1)) \\ &\qquad - d(t_j + 1)(x^*(t_j + 1) - x(t_j + 1))) + p(t)x(t + 1) \\ &= \sum_{j=k}^{l} p(t_j + 1)(f(x^*(t_j + 1)) - f(x_*) - (x^*(t_j + 1) - x_*)) \\ &\qquad + \sum_{j=k}^{l} p(t_j + 1)(f(x_*) - x_* - (f(x(t_j + 1)) - d(t_j + 1)x(t_j + 1))) \\ &\qquad + \sum_{j=k}^{l} p(t_j + 1)x^*(t_j + 1)(1 - d(t_j + 1)) + p(t)x(t + 1) \\ &\leqq (l - k + 1)LN - (l - k + 1)LN/2 + p(r)K < e. \end{aligned}$$

4. A characterization of optimal consumption plans. We consider a one-good economy which is described by a production function f and a utility function u. $f(x)$ is assumed to be a concave function which is defined for $x \geqq 0$ and which satisfies assumptions (i)–(iv) of § 2. $u(y)$ is assumed to be a concave and increasing function which is defined for $y \geqq 0$ (the case $u(0) = -\infty$ is not excluded). In addition, we assume that the initial capital stock is positive, i.e., $x_0 > 0$.

DEFINITION 4.1. A feasible consumption plan $\underset{\sim}{y}$ is *optimal* if

$$\limsup \sum_{s=0}^{t} (u(y^*(s)) - u(y(s))) \leqq 0 \qquad \text{for all } \underset{\sim}{y}^* \in Y.$$

DEFINITION 4.2. A feasible program $(\underset{\sim}{x}, \underset{\sim}{y})$ is *competitive* if there exists a sequence $\underset{\sim}{p} > \underset{\sim}{0}$ such that:

(i) $\underset{\sim}{p}$ is an i.p.m. price system for $\underset{\sim}{x}$;

(ii) $u(y(t)) - p(t)y(t) \geqq u(y) - p(t)y$ for all $y \geqq 0$.

$\underset{\sim}{p}$ will be called a *system of competitive prices* for $(\underset{\sim}{x}, \underset{\sim}{y})$.

LEMMA 4.3. *There exists a feasible program* $(\underset{\sim}{x}^*, \underset{\sim}{y}^*)$ *with the following properties:* $\underset{\sim}{y}^* \gg \underset{\sim}{0}$, *and there exists a natural number* t_0 *such that* $y^*(t) = y_*$ *for* $t > t_0$.

Proof. If $x_0 \geqq x_*$, let $x^*(t) = x_*$ for $t \geqq 1$. If $\underset{\sim}{y}^*$ is the feasible consumption plan which is determined by $\underset{\sim}{x}^*$, then it has the desired properties. If $x_0 < x_*$, then $f(x_0) \geqq f'_+(x_0)x_0 > x_0$. There exist $e > 0$ and $0 < w < 1$ such that $f(x_* + e) > x_* + e$ and

(a) $f(x_0) > wx_0 + (1 - w)(x_* + e)$.

Define a sequence $\underset{\sim}{x}$ by $x(0) = x_0$ and

(b) $x(t) = wx(t - 1) + (1 - w)(x_* + e)$, $t = 1, 2, \cdots$.

Let $\underset{\sim}{y}$ be defined by

$$y(t) = f(x(t)) - x(t + 1), \qquad t = 0, 1, 2, \cdots.$$

We claim that $\underset{\sim}{y} \gg \underset{\sim}{0}$. By (a) and (b), $y(0) > 0$. Assume that $y(t) > 0$, $t \geqq 0$. Then

$$\begin{aligned} y(t + 1) &= f(x(t + 1)) - x(t + 2) \\ &= f(wx(t) + (1 - w)(x_* + e)) - wx(t + 1) - (1 - w)(x_* + e) \\ &\geqq w(f(x(t)) - x(t + 1)) + (1 - w)(f(x_* + e) - (x_* + e)) > 0. \end{aligned}$$

It follows from (b) that $x(t) = w^t x_0 + (1 - w^t)(x_* + e)$, $t = 0, 1, 2, \cdots$. Hence, there exists a natural number t_0 such that $x(t) > x_*$ for $t > t_0$. Let $\underset{\sim}{x}^*$ be defined by $x^*(t) = x(t)$ for $t \leqq t_0$, and $x^*(t) = x_*$ for $t > t_0$. Then $\underset{\sim}{x}^*$ is a feasible capital sequence. Let $\underset{\sim}{y}^*$ be the feasible consumption plan which is determined by $\underset{\sim}{x}^*$. Since $\underset{\sim}{x} > \underset{\sim}{x}^*$, $\underset{\sim}{y}^* > \underset{\sim}{y}$. Hence, $\underset{\sim}{y}^* \gg \underset{\sim}{0}$. In addition $y^*(t) = y_*$ for $t > t_0$.

LEMMA 4.4. *Let $\underset{\sim}{y}$ be an optimal consumption plan. There exists a real number M such that*

$$\sum_{s=0}^{t} (u(y_*) - u(y(s))) \leqq M, \qquad t = 0, 1, 2, \cdots. \tag{4.1}$$

Proof. By Lemma 4.3 there exists a feasible consumption plan $\underset{\sim}{y}^*$ such that $\underset{\sim}{y}^* \gg \underset{\sim}{0}$, and there exists a natural number t_0 such that $y^*(t) = y_*$ for $t > t_0$. There exists a natural number $r \geqq t_0$ such that if $t > r$, then $\sum_{s=0}^{t} (u(y^*(s)) - u(y(s))) \leqq 1$. Since $\underset{\sim}{y}^* \gg \underset{\sim}{0}$, $u(y(s)) > -\infty$, $s = 0, 1, 2, \cdots$. For $t > r$,

$$\begin{aligned} \sum_{s=0}^{t} (u(y_*) - u(y(s))) &= \sum_{s=0}^{r} (u(y_*) - u(y^*(s))) + \sum_{s=0}^{t} (u(y^*(s)) - u(y(s))) \\ &\leqq \sum_{s=0}^{r} (u(y_*) - u(y^*(s))) + 1. \end{aligned}$$

Thus, the existence of a real number M which satisfies (4.1) is now clear.

LEMMA 4.5. *Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program and assume that $\underset{\sim}{y}$ is an optimal consumption plan. Then* $\lim x(t) = x_*$.

Proof. There exists a real number $p > 0$ such that

$$u(y_*) - py_* \geqq u(y) - py \qquad \text{for all } y \geqq 0.$$

For $t \geqq 0$ we have

$$\sum_{s=0}^{t} (u(y_*) - u(y(s))) \geqq p \sum_{s=0}^{t} (y_* - y(s))$$

$$= p(f(x_*) - f(x_0)) \tag{4.2}$$

$$+ p \sum_{s=1}^{t} (f(x_*) - x_* - (f(x(s)) - x(s))) + p(x(t+1) - x_*).$$

We shall now prove:

(a) $\liminf x(t) \geqq x_*$.

Assume, on the contrary, that $\liminf x(t) = z < x_*$. There exists an $L > 0$ such that $f(x_*) - x_* - (f(x) - x) \geqq L/p$ for $0 \leqq x \leqq (z + x_*)/2$. There exists an increasing sequence of natural numbers $\{t_1, t_2, \cdots, t_j, \cdots\}$ such that $x(t_j) \leqq (z + x_*)/2, j = 1, 2, \cdots$. If $t \geqq t_k$, then it follows from (4.2) that

$$\sum_{s=0}^{t} (u(y_*) - u(y(s))) \geqq -p(f(x_0) + x_*) + (k-1)L,$$

which contradicts Lemma 4.4.

We shall now prove:

(b) $\limsup x(t) \leqq x_*$.

Assume, on the contrary, that $\limsup x(t) = z > x_*$. By Lemma 3.6, $\underset{\sim}{x}$ is bounded; hence, $z < \infty$. We now distinguish the following cases.

(b.1) $f(z) - z < f(x_*) - x_*$.

There exist $e > 0$ and $L > 0$ such that

$$f(x_*) - x_* - (f(x) - x) \geqq L/p \qquad \text{for } x \geqq z - e.$$

There exists an increasing sequence of natural numbers $\{t_1, t_2, \cdots, t_j, \cdots\}$ such that $x(t_j) \geqq z - e, j = 1, 2, \cdots$. If $t \geqq t_k$, then it follows that (4.2) that

$$\sum_{s=0}^{t} (u(y_*) - u(y(s))) \geqq -p(f(x_0) + x_*) + (k-1)L,$$

which contradicts Lemma 4.4.

(b.2) $f(z) - z = f(x_*) - x_*$.

It follows from the concavity of f that

$$f(x) - x = f(x_*) - x_* \qquad \text{for } x_* \leqq x \leqq z. \tag{4.3}$$

We now distinguish the following possibilities.

(b.2.1) $\liminf x(t) = x_*$.

There exists a natural number $s \geqq 1$ such that $f(x(s-1)) > x_*$ and

$$|u(f(x(s-1)) - x_*) - u(y(s-1))| + p|f(x(s)) - f(x_*)| < p(z - x_*)/2.$$

Define a feasible capital sequence $\underset{\sim}{x}^*$ by $x^*(t) = x(t), t < s$, and $x^*(t) = x_*, t \geqq s$.

Let $\underset{\sim}{y}^*$ be the feasible consumption plan which is determined by $\underset{\sim}{x}^*$. For $t > s$,

$$\begin{aligned}
&\sum_{r=0}^{t} (u(y^*(r)) - u(y(r))) \\
&\quad = u(f(x(s-1)) - x_*) - u(y(s-1)) + \sum_{r=s}^{t} (u(y_*) - u(y(r))) \\
&\quad \geqq u(f(x(s-1)) - x_*) - u(y(s-1)) + p(f(x_*) - f(x(s))) \\
&\qquad + p \sum_{r=s+1}^{t} (f(x_*) - x_* - (f(x(r)) - x(r))) + p(x(t+1) - x_*) \\
&\quad > -p(z - x_*)/2 + p(x(t+1) - x_*).
\end{aligned}$$

There exists an increasing sequence of natural numbers $\{t_1, t_2, \cdots, t_j, \cdots\}$ such that $t_1 > s$ and $x(t_j + 1) > x_*/4 + 3z/4, j = 1, 2, \cdots$.

$$\begin{aligned}
\sum_{r=0}^{t_j} (u(y^*(r)) - u(y(r))) &> -p(z - x_*)/2 + p(x(t_j + 1) - x_*) \\
&> -p(z - x_*)/2 + 3p(z - x_*)/4 \\
&= p(z - x_*)/4, \qquad j = 1, 2, \cdots,
\end{aligned}$$

which contradicts our assumption that $\underset{\sim}{y}$ is optimal.

(b.2.2) $\liminf x(t) > x_*$.

There exist an $e > 0$ and a natural number t_0 such that

$$x_* + e \leqq x(t) \leqq z + e \qquad \text{for } t \geqq t_0.$$

Hence,

$$x_* \leqq x(t) - e \leqq z \qquad \text{for } t \geqq t_0. \tag{4.4}$$

Define a sequence $\underset{\sim}{x}^*$ by $x^*(t) = x(t)$, $t \leqq t_0$, and $x^*(t) = x(t) - e$, $t > t_0$. Let $\underset{\sim}{y}^*$ be defined by

$$y^*(t) = f(x^*(t)) - x^*(t+1), \qquad t = 0, 1, 2, \cdots.$$

If $t < t_0$, then $y^*(t) = y(t)$. $y^*(t_0) = f(x(t_0)) - (x(t_0 + 1) - e) = y(t_0) + e$. For $t > t_0$, it follows from (4.3) and (4.4) that

$$\begin{aligned}
y^*(t) - y(t) &= f(x(t) - e) - (x(t+1) - e) - (f(x(t)) - x(t+1)) \\
&= f(x(t) - e) - (x(t) - e) - (f(x(t)) - x(t)) \geqq 0.
\end{aligned}$$

Thus $\underset{\sim}{y}^* > \underset{\sim}{y}$. Hence, $\underset{\sim}{y}^*$ is feasible. Since u is increasing, the inequality $\underset{\sim}{y}^* > \underset{\sim}{y}$ contradicts the optimality of $\underset{\sim}{y}$.

THEOREM 4.6. *Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a feasible program and assume that $\underset{\sim}{y}$ is an optimal consumption plan. Then $(\underset{\sim}{x}, \underset{\sim}{y})$ is competitive, and for any competitive price system $\underset{\sim}{p}$ of $(\underset{\sim}{x}, \underset{\sim}{y})$ the sequence q, where $q(t) = p(t)x(t+1)$, $t = 0, 1, 2, \cdots$, is bounded.*

Proof. By a theorem of Gale and Sutherland, $(\underset{\sim}{x}, \underset{\sim}{y})$ is competitive [2, Thm. 4]. Let $\underset{\sim}{p}$ be a system of competitive prices for $(\underset{\sim}{x}, \underset{\sim}{y})$. By Lemma 4.5, $\lim x(t) = x_*$.

Hence, $\lim y(t) = y_*$. Thus, there exists a natural number t_0 such that $y(t) \geqq y_*/2 > 0$, $t > t_0$. It follows from (ii) of Definition 4.2 that

$$p(t) \leqq u'_-(y(t)) \leqq u'_-(y_*/2), \qquad t > t_0.$$

Hence, p is bounded. By Lemma 3.10, q is bounded.

LEMMA 4.7. *Let (x, y) be a competitive program. Assume that (x, y) has a bounded system p of competitive prices, and that there exists a natural number r such that $x(t) = x_*$ for $t \geqq r$. Then y is an optimal consumption plan.*

Proof. Let (x^*, y^*) be any feasible program. We distinguish the following cases.

(a) $\liminf x^*(t) < x_*$.

Let $t \geqq r$. Then

$$\sum_{s=0}^{t} (u(y(s)) - u(y^*(s))) = \sum_{s=0}^{r-1} (u(y(s)) - u(y^*(s))) + \sum_{s=r}^{t} (u(y_*) - u(y^*(s))).$$

It follows from (ii) of Definition 4.2 that $u(y(s)) > -\infty$, $s = 0, 1, 2, \cdots$. Hence,

$$\sum_{s=0}^{r-1} (u(y(s)) - u(y^*(s))) > -\infty.$$

There exists a real number $p > 0$ such that

$$u(y_*) - py_* \geqq u(y) - py \qquad \text{for all } y \geqq 0.$$

Thus

$$\begin{aligned}\sum_{s=r}^{t} (u(y_*) - u(y^*(s))) &\geqq p \sum_{s=r}^{t} (y_* - y^*(s)) \\ &= p(f(x_*) - f(x^*(r))) \\ &\quad + p \sum_{s=r+1}^{t} (f(x_*) - x_* - (f(x^*(s)) - x^*(s))) \\ &\quad + p(x^*(t+1) - x_*).\end{aligned}$$

There exist an increasing sequence of natural numbers $\{t_1, t_2, \cdots, t_j, \cdots\}$ and a real number $e > 0$ such that $t_1 \geqq r + 1$, and $x^*(t_j) \leqq x_* - e$, $j = 1, 2, \cdots$. There exists a real number $L > 0$ such that

$$f(x_*) - x_* - (f(x) - x) \geqq L/p \qquad \text{for } 0 \leqq x \leqq x_* - e.$$

If $t \geqq t_k$, then

$$\sum_{s=0}^{t} (u(y(s)) - u(y^*(s))) \geqq \sum_{s=0}^{r-1} (u(y(s)) - u(y^*(s))) - p(f(x^*(r)) + x_*) + kL.$$

Hence, $\lim \sum_{s=0}^{t} (u(y(s)) - u(y^*(s))) = \infty$.

(b) $\liminf x^*(t) \geqq x_*$.

Let $t \geqq r$. Then

$$\sum_{s=0}^{t} (u(y(s)) - u(y^*(s))) = \sum_{s=0}^{t} (u(y(s)) - p(s)(y(s) - f(x(s)) + x(s+1)))$$

$$- \sum_{s=0}^{t} (u(y^*(s)) - p(s)(y^*(s) - f(x^*(s)) + x^*(s+1)))$$

$$= \sum_{s=0}^{t} (u(y(s)) - p(s)y(s) - (u(y^*(s)) - p(s)y^*(s)))$$

$$+ \sum_{s=1}^{t} (p(s)f(x(s)) - p(s-1)x(s) - (p(s)f(x^*(s)) - p(s-1)x^*(s)))$$

$$+ p(t)(x^*(t+1) - x_*)$$

$$\geqq p(t)(x^*(t+1) - x_*).$$

By assumption, there exists a real number R such that $p(t) \leqq R$, $t = 0, 1, 2, \cdots$. Let $e > 0$ be given. It follows from (b) that there exists a natural number $t_0 \geqq r$ such that $x^*(t+1) - x_* > -e/R$ for $t \geqq t_0$. Hence, for $t \geqq t_0$,

$$\sum_{s=0}^{t} (u(y(s)) - u(y^*(s))) \geqq p(t)(x^*(t+1) - x_*) > -e.$$

LEMMA 4.8. *Let $(\underset{\sim}{x}, \underset{\sim}{y})$ be a competitive program and let $\underset{\sim}{p}$ be a system of competitive prices for $(\underset{\sim}{x}, \underset{\sim}{y})$. Then there exists a real number $M > 0$ such that $p(t) \geqq M$, $t \geqq 0$, and*

$$(p(t+1) - p(t))(y(t+1) - y(t)) \leqq 0, \qquad t = 0, 1, 2, \cdots. \tag{4.5}$$

Proof. By (ii) of Definition 4.2, $p(t)$ is a subderivative of u at $y(t)$, $t \geqq 0$. It follows from Lemma 3.6 that there exists a real number N such that $y(t) \leqq N$, $t \geqq 0$. Hence, $p(t) \geqq u'_+(N) > 0$, $t \geqq 0$. To prove (4.5), observe that

$$u(y(t)) - u(y(t+1)) \geqq p(t)(y(t) - y(t+1)),$$

and

$$u(y(t+1)) - u(y(t)) \geqq p(t+1)(y(t+1) - y(t)).$$

Adding the last two inequalities, we get (4.5).

We now add the following assumption:

$$f(x) - x < f(x_*) - x_* \qquad \text{for } x > x_*. \tag{4.6}$$

It follows from (4.6) that

$$f'_-(x) < 1 \qquad \text{for } x > x_*. \tag{4.7}$$

LEMMA 4.9. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a competitive program. If there exists a natural number* t *such that*

$$x(t+1) > x_* \quad \text{and} \quad x(t+1) \geqq x(t),$$

then

$$x(t+h) \geqq x(t+h-1) \qquad \text{for } h = 1, 2, \cdots.$$

Proof. The proof is by induction on h. Assume for $h \geqq 1$ that

$$x(t+l) \geqq x(t+l-1), \qquad l = 1, \cdots, h.$$

Let $\underset{\sim}{p}$ be a system of competitive prices for $(\underset{\sim}{x}, \underset{\sim}{y})$. By Lemma 4.8, $\underset{\sim}{p} \gg \underset{\sim}{0}$. Hence, by (ii) of Lemma 3.2, $p(t+h-1)/p(t+h)$ is a subderivative of f at $x(t+h)$. $x(t+h) \geqq x(t+1) > x_*$. Hence, by (4.7), $f'_-(x(t+h)) < 1$. Hence, $p(t+h-1) < p(t+h)$. It follows now from (4.5) that $y(t+h-1) \geqq y(t+h)$. Thus,

$$\begin{aligned} x(t+h+1) - x(t+h) &= f(x(t+h)) - f(x(t+h-1)) + y(t+h-1) - y(t+h) \\ &\leqq 0. \end{aligned}$$

LEMMA 4.10. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a competitive program. If there exists a natural number* t *such that*

$$x(t+1) < x_* \quad \text{and} \quad x(t+1) \leqq x(t),$$

then

$$x(t+h) \leqq x(t+h-1) \qquad \text{for } h = 1, 2, \cdots.$$

The proof, which is similar to the proof of Lemma 4.9, is omitted.

LEMMA 4.11. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a competitive program. If there exists a natural number* t *such that* $x(t) \geqq x_*$, *then* $x(s) \geqq x_*$ *for* $s \geqq t$.

Proof. Assume, on the contrary, that there exists an $s \geqq t$ such that $x(s) < x_*$. Let

$$r = \min \{s | s \geqq t \text{ and } x(s) < x_*\}.$$

Then, $x(r) < x_*$ and $x(r) < x(r-1)$. By Lemma 4.10, $x(r) \geqq x(r+h)$, $h \geqq 1$. Let $\underset{\sim}{p}$ be a system of competitive prices for $(\underset{\sim}{x}, \underset{\sim}{y})$, $p(r+h-1)/p(r+h) \geqq f'_+(x(r)) > 1$, $h \geqq 1$. Hence, $\lim p(s) = 0$, which contradicts Lemma 4.8.

LEMMA 4.12. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a competitive program. Assume that* $(\underset{\sim}{x}, \underset{\sim}{y})$ *has a bounded system* $\underset{\sim}{p}$ *of competitive prices. Then, if there exists a natural number* t *such that* $x(t) \leqq x_*$, *then* $x(s) \leqq x_*$ *for* $s \geqq t$.

Proof. Assume, on the contrary, that there exists an $s \geqq t$ such that $x(s) > x_*$. Let

$$r = \min \{s | s \geqq t \text{ and } x(s) > x_*\}.$$

Then $x(r) > x_*$ and $x(r) > x(r-1)$. By Lemma 4.9, $x(r+h) \geqq x(r)$ for $h \geqq 1$. By (4.7), $f'_-(x(r)) < 1$. Hence, $p(r+h-1)/p(r+h) \leqq f'_-(x(r)) < 1$, $h \geqq 1$. Thus, $\lim p(s) = \infty$, which contradicts our assumption that $\underset{\sim}{p}$ is bounded.

COROLLARY 4.13. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a competitive program. Assume that* $(\underset{\sim}{x}, \underset{\sim}{y})$ *has a bounded system of competitive prices. Then, if there exists a natural number* t *such that* $x(t) = x_*$, *then* $x(s) = x_*$ *for* $s \geqq t$.

Proof. This follows from Lemma 4.11 and Lemma 4.12.

LEMMA 4.14. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a competitive program. Assume that* $(\underset{\sim}{x}, \underset{\sim}{y})$ *has a bounded system* $\underset{\sim}{p}$ *of competitive prices. If* $x_0 < x_*$, *then either there exists a* t *such that* $x(s) = x_*$ *for* $s \geqq t$ *or* $x(s) < x_*$ *for* $s \geqq 0$ *and* $\lim x(s) = x_*$.

Proof. $x(0) = x_0 < x_*$. By Lemma 4.12, $x(t) \leqq x_*$ for $t \geqq 0$. If there exists a t such that $x(t) = x_*$, then, by Corollary 4.13, $x(s) = x_*$ for $s \geqq t$. Otherwise, $x(t) < x_*$ for $t \geqq 0$. Hence, $p(t+1) < p(t)$, $t = 0, 1, 2, \cdots$. If $\liminf x(t) < x_*$, then $\lim p(t) = 0$, which contradicts Lemma 4.8.

LEMMA 4.15. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a competitive program. Assume that* $(\underset{\sim}{x}, \underset{\sim}{y})$ *has a bounded system* $\underset{\sim}{p}$ *of competitive prices. If* $x_0 > x_*$, *then either there exists a* t *such that* $x(s) = x_*$ *for* $s \geqq t$ *or* $x(s) > x_*$ *for* $s \geqq 0$ *and* $\lim x(s) = x_*$.

Proof. $x(0) = x_0 > x_*$. By Lemma 4.11, $x(t) \geqq x_*$ for $t \geqq 0$. If there exists a t such that $x(t) = x_*$, then, by Corollary 4.13, $x(s) = x_*$ for $s \geqq t$. Otherwise, $x(t) > x_*$ for $t \geqq 0$. Hence, it follows from (4.7) that $p(t+1) > p(t)$, $t = 0, 1, 2, \cdots$. If $\limsup x(t) > x_*$, then it follows from (4.7) that $\lim p(t) = \infty$, which contradicts our assumption that p is bounded.

THEOREM 4.16. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a competitive program and let* $\underset{\sim}{p}$ *be a system of competitive prices for* $(\underset{\sim}{x}, \underset{\sim}{y})$. *If there exists a real number* R *such that* $p(t)x(t+1) \leqq R$, $t = 0, 1, 2, \cdots$, *then* $\underset{\sim}{y}$ *is an optimal consumption plan.*

Proof. By Lemma 3.10, $\underset{\sim}{p}$ is bounded. Hence, by Corollary 4.13, Lemma 4.14 and Lemma 4.15, exactly one of the following possibilities holds:

(i) There exists a t such that $x(s) = x_*$ for $s \geqq t$.

(ii) $x(t) > x_*$ for $t \geqq 0$ and $\lim x(t) = x_*$.

(iii) $x(t) < x_*$ for $t \geqq 0$ and $\lim x(t) = x_*$.

If (i) holds, then by Lemma 4.7, $\underset{\sim}{y}$ is an optimal consumption plan. If (ii) holds, then by Lemma 3.14, $\underset{\sim}{p}$ is a system of efficiency prices of type V for $\underset{\sim}{y}$. Let $(\underset{\sim}{x}^*, \underset{\sim}{y}^*)$ be any feasible program. By (ii) of Definition 4.2,

$$\sum_{s=0}^{t} (u(y^*(s)) - u(y(s))) \leqq \sum_{s=0}^{t} p(s)(y^*(s) - y(s)).$$

Hence,

$$\limsup \sum_{s=0}^{t} (u(y^*(s)) - u(y(s))) \leqq \limsup \sum_{s=0}^{t} p(s)(y^*(s) - y(s)) \leqq 0.$$

If (iii) holds, then, by Lemma 3.15, $\underset{\sim}{p}$ is a system of efficiency prices of type V for $\underset{\sim}{y}$, and again this implies that $\underset{\sim}{y}$ is an optimal consumption plan.

Example 4.17. Let $f(x) = \min(2x, 1 + x, 2 + x/2)$ and let $u(y) = 2y^{1/2}$. Let $x(t) = 2$, $y(t) = 1$, and $p(t) = 1$, $t = 0, 1, 2, \cdots$. Then $(\underset{\sim}{x}, \underset{\sim}{y})$ is a competitive program and $\underset{\sim}{p}$ is a system of competitive prices for it. However, $\underset{\sim}{y}$ is not an optimal consumption plan.

We conclude from Example 4.17 that assumption (4.6) is essential for the validity of Theorem 4.16.

We now drop assumption (4.6).

COROLLARY 4.18. *Let* $(\underset{\sim}{x}, \underset{\sim}{y})$ *be a feasible program and assume that* $\underset{\sim}{y}$ *is an optimal consumption plan. Then* $\underset{\sim}{y}$ *has a system of efficiency prices of type* V.

Proof. By Theorem 4.6, $(\underset{\sim}{x}, \underset{\sim}{y})$ has bounded system $\underset{\sim}{p}$ of competitive prices. We now distinguish the following cases.

(a) Assumption (4.6) holds, i.e., $f(x) - x < f(x_*) - x_*$ for $x > x_*$.

Then, exactly one of the possibilities (i)–(iii) in the proof of Theorem 4.16 holds. If (i) holds, then let p^* be defined by $p^*(s) = p(s)$ for $s \leqq t$, and $p^*(s) = p(t)$ for $s > t$. Then, by Lemma 3.15, p^* is a system of efficiency prices of type V for y. If (ii) holds, then, by Lemma 3.14, p is a system of efficiency prices of type V for y. If (iii) holds, then, by Lemma 3.15, p is a system of efficiency prices of type V for y.

(b) There exists a $z > x_*$ such that $f(z) - z = f(x_*) - x_*$.

Then, $f(x) - x = f(x_*) - x_*$ for $x_* \leqq x \leqq z$. Hence, $f'_+(x) \geqq 1$ for $x < z$. By Lemma 4.5, $\lim x(t) = x_*$. Hence, there exists an r such that $x(t) < z$ for $t \geqq r$. For $t \geqq r$, $p(t)/p(t+1) \geqq f'_+(x(t+1)) \geqq 1$. Hence, by Lemma 3.15, p is a system of efficiency prices of type V for y.

Remark 4.19. It is known [1, §4], that not every efficient consumption plan has a system of efficiency prices of type V.

Acknowledgment. I am grateful to M. Yaari for many helpful discussions.

REFERENCES

[1] D. Cass and M. E. Yaari, *Present values playing the role of efficiency prices in the one-good growth model*, Research Rep. 14, Department of Economics, The Hebrew University of Jerusalem, 1970.

[2] D. Gale and W. R. Sutherland, *Analysis of a one-good model of economic development*, Mathematics of the Decision Sciences, Part 2, vol. 12, G. B. Dantzig and A. F. Veinott, Jr., eds., Lectures in Appl. Math., American Mathematical Society, Providence, R.I., 1968, pp. 120–136.

[3] B. Peleg and M. E. Yaari, *Efficiency prices in an infinite-dimensional space*, J. Economic Theory, 2 (1970), pp. 41–85.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

A GEOMETRICALLY CONVERGENT ALGORITHM FOR SOLVING OPTIMAL CONTROL PROBLEMS*

EARL R. BARNES†

Abstract. Generalized versions of the Frank–Wolfe algorithm have been used by several authors to solve problems in optimal control (cf. [1], [8], [9], [10] and [11]). All of these algorithms involve a linearization of the given problem at each iteration. In the present paper, we offer a modification of these algorithms which does not require the problem to be completely linearized at each iteration. The result is that we obtain a geometrically convergent algorithm. It is known that the Frank–Wolfe algorithm cannot converge geometrically fast unless some restrictions are placed on the cost functional and the constraint set (cf. [3] and [4, p. 24]). We present some numerical results to show that the difference in performances of the two algorithms can be quite significant.

1. Introduction. In 1966 Gilbert [1] described an algorithm for solving a certain class of optimal control problems. This algorithm generalized the Frank–Wolfe algorithm for quadratic programming [2]. Appearing shortly after Gilbert's paper was a paper by Canon and Cullum [3] showing that the Frank–Wolfe algorithm can actually exhibit a very poor rate of convergence. Barr observed this same type of behavior in Gilbert's algorithm and proposed a modification which substantially increases the rate of convergence. This work was first reported in [5] and subsequently in [6] and [7].

In [8], the present author described a broad class of problems to which Gilbert's algorithm, extended to infinite-dimensional spaces, applies. A few of these problems had already been solved by generalized versions of the Frank–Wolfe algorithm in [9]–[11]. Unfortunately, Barr's modification of Gilbert's algorithm is not applicable to any of these problems, except the special case of Gilbert's original problem. This situation is remedied to a certain extent in the present paper. Here, we describe an algorithm which is applicable to most of the problems in [8] and which exhibits a geometric rate of convergence.

For problems where the constraint set is strongly convex, and where the gradient of the cost functional is bounded away from zero on the constraint set, it is known that the Frank–Wolfe algorithm will also converge geometrically fast (cf. [4, p. 24]). However, this situation is violated quite often in practice. For example, in problems such as the one we discuss in § 3, the constraint set is not strongly convex, and the Frank–Wolfe algorithm cannot converge geometrically fast for such an example. See the note in [4, p. 26]. Another such example is discussed in [3]. In problems where the optimum lies interior to the constraint set, the gradient of the cost functional vanishes at the optimum and again the condition for geometric convergence of the Frank–Wolfe algorithm is violated. No such conditions are required for the geometric convergence of our method. The amount of work involved in implementing one step of each algorithm is approximately the same.

2. The optimal control problem. Let $g(x)$, $f(t, x)$ and $h(t, u)$ be real-valued continuous functions defined for $t \in [0, T]$ and for all $x \in E_n$ and $u \in E_m$. Assume

* Received by the editors October 27, 1970, and in final revised form August 11, 1971.

† Thomas J. Watson Research Center, International Business Machines Corporation, Yorktown Heights, New York 10598.

further that these functions are twice continuously differentiable and convex with respect to the variables x and u. We further restrict h to be a quadratic form in u. Denote by ∇f and $\nabla^2 f$, respectively, the gradient and Hessian of f with respect to x. The gradients and Hessians of g and h with respect to x and u will be denoted similarly. We assume that the Hessian $\nabla^2 h(t)$ (which is independent of u) is positive definite for each $t \in [0, T]$.

Consider the optimal control problem of minimizing the cost functional

$$c(u) = g(x(T)) + \int_0^T f(t, x(t)) + h(t, u(t))\, dt \tag{2.1}$$

subject to the differential equation constraint

$$\dot{x} = A(t)x + B(t)u, \quad 0 \leqq t \leqq T, \quad x(0) = x_0 \tag{2.2}$$

and the constraint $u \in \mathscr{C}$ on the function u. A and B are continuous matrices of dimensions $n \times n$ and $n \times m$, respectively. $x_0 \in E_n$ is fixed. $\mathscr{C}$ denotes the class of admissible controls defined as follows.

DEFINITION 2.1. Let $\Omega \subset E_m$ be a given compact convex constraint set. A measurable function u defined on $[0, T]$ and taking values in Ω is said to be an *admissible control.*

DEFINITION 2.2. An admissible control which minimizes the functional (2.1) on the class of admissible controls is said to be an *optimal control.*

The existence of an optimal control has been established by Cesari [12] and by Lee and Markus in [13, Chap. 3].

A sequence of admissible controls converging to the optimal control can be generated by the following iterative procedure.

(i) Let $u_1 \in \mathscr{C}$ be chosen arbitrarily.

(ii) If $u_1, \cdots, u_k \in \mathscr{C}$ have been chosen, choose $\tilde{u}_k \in \mathscr{C}$ by the requirement that the minimum

$$\min_{u \in \mathscr{C}} x(T) \cdot \nabla g(x_k(T)) + \int_0^T x \cdot \nabla f(t, x_k) + h(t, u)\, dt$$

be attained at $u = \tilde{u}_k$. Here x and x_k are the responses of the system (1.2) due to u and u_k, respectively. Similarly, we shall denote by $\tilde{x}_k$ the response due to $\tilde{u}_k$. Pontryagin's maximum principle provides a recipe for computing $\tilde{u}_k$ (cf. [14]).

(iii) Let $u_{k+1} = u_k + \alpha_k(\tilde{u}_k - u_k)$, where $\alpha_k \in [0, 1]$ is defined by the requirement that

$$c(u_{k+1}) = \min_{0 \leqq \alpha \leqq 1} c(u_k + \alpha(\tilde{u}_k - u_k)).$$

To facilitate our discussion of this algorithm, we adopt a shorthand notation. For any positive semidefinite matrices $C(t)$, G and $D(t)$ of dimension $n \times n$, $n \times n$, and $m \times m$, respectively, we shall write

$$\|x\|_C^2 = \int_0^T x(t) \cdot C(t)x(t)\, dt, \qquad \|u\|_D^2 = \int_0^T u(t) \cdot D(t)u(t)\, dt$$

and

$$|x(T)|_G^2 = x(T) \cdot Gx(T).$$

The Euclidean norm of a matrix A defined on E_n (or E_m) will be denoted by $|A|_n$ (or $|A|_m$). In many cases, we shall suppress the dependence of functions and

matrices on the argument t. This has been done already in (ii) for the functions x, x_k and u.

THEOREM 2.1. *Let u^* denote an optimal control for the system* (2.1)–(2.2). *Then the sequence $\{u_k\}$ generated in* (i), (ii) *and* (iii) *converges pointwise to u^* for almost all $t \in [0,T]$.*

Before proving this theorem, we wish to point out that since h is quadratic in u, the function $\tilde{u}_k$ in (ii) can be determined by minimizing the functional

$$x(T)\cdot\nabla g(x_k(T)) + \int_0^T \{\nabla f(t,x_k)\cdot x + \nabla h(t,u_k)\cdot u + \tfrac{1}{2}(u-u_k)\cdot\nabla^2 h(t)(u-u_k)\}\,dt. \tag{2.3}$$

This is the property of $\tilde{u}_k$ that is used in the proof of the theorem.

Proof of Theorem 1.2. Let $m = c(u^*)$. Then by writing $u^* = u_k + (u^* - u_k)$ and expanding c about u_k, we obtain

$$\begin{aligned} m = c(u^*) \geqq c(u_k) &+ (x^*(T) - x_k(T))\cdot\nabla g(x_k(T)) \\ &+ \int_0^T \{(x^* - x_k)\cdot\nabla f(t,x_k) + (u^* - u_k)\cdot\nabla h(t,u_k) \\ &\quad + \tfrac{1}{2}(u^* - u_k)\cdot\nabla^2 h(t)(u^* - u_k)\}\,dt. \end{aligned} \tag{2.4}$$

Here we have used the convexity properties of g and f. x^* denotes the optimal trajectory corresponding to u^*. Since $\tilde{u}_k$ minimizes the functional (2.3) over the set $\mathscr{C}$, the right-hand side of (2.4) cannot increase if we replace u^* and x^* by $\tilde{u}_k$ and $\tilde{x}_k$ respectively. Therefore

$$\begin{aligned} m - c(u_k) \geqq (\tilde{x}_k(T) - x_k(T))&\cdot\nabla g(x_k(T)) \\ &+ \int_0^T \{(\tilde{x}_k - x_k)\cdot\nabla f(t,x_k) + (\tilde{u}_k - u_k)\cdot\nabla h(t,u_k) \\ &\quad + \tfrac{1}{2}(\tilde{u}_k - u_k)\cdot\nabla^2 h(t)(\tilde{u}_k - u_k)\}\,dt. \end{aligned} \tag{2.5}$$

Let $\alpha \in [0,1]$ be fixed and let $u = u_k + \alpha(\tilde{u}_k - u_k)$. Then by Taylor's theorem,

$$\begin{aligned} c(u) - m = c(u_k) - m &+ \alpha(\tilde{x}_k(T) - x_k(T))\cdot\nabla g(x_k(T)) \\ &+ \alpha\int_0^T \{\nabla f(t,x_k)\cdot(\tilde{x}_k - x_k) + \nabla h(t,u_k)\cdot(\tilde{u}_k - u_k) \\ &\quad + \tfrac{1}{2}(\tilde{u}_k - u_k)\cdot\nabla^2 h(t)(\tilde{u}_k - u_k)\}\,dt \\ &+ \frac{\alpha^2}{2}(\tilde{x}_k(T) - x_k(T))\cdot\nabla^2 g(\xi)(\tilde{x}_k(T) - x_k(T)) \\ &+ \frac{\alpha^2}{2}\int_0^T \{(\tilde{x}_k - x_k)\cdot\nabla^2 f(t,\zeta(t))(\tilde{x}_k - x_k) + (\tilde{u}_k - u_k) \\ &\quad \cdot\nabla^2 h(t)(\tilde{u}_k - u_k)\}\,dt \\ &- \frac{\alpha}{2}\int_0^T (\tilde{u}_k - u_k)\cdot\nabla^2 h(t)(\tilde{u}_k - u_k)\,dt. \end{aligned}$$

(ξ and $\zeta(t)$ are intermediate values determined by Taylor's theorem.) If we apply (2.5) to the right-hand side of this equation and make use of the shorthand notation introduced above, we obtain

$$\begin{aligned} c(u) - m &\leqq c(u_k) - m + \alpha(m - c(u_k)) + \varphi(\alpha) \\ &= (1 - \alpha)(c(u_k) - m) + \varphi(\alpha), \end{aligned} \tag{2.6}$$

where

$$\varphi(\alpha) = \tfrac{1}{2}\{(\alpha^2 - \alpha)\|\tilde{u}_k - u_k\|^2_{\nabla^2 h} + \alpha^2\|\tilde{x}_k - x_k\|^2_{\nabla^2 f(t,\zeta)} + \alpha^2|\tilde{x}_k(T) - x_k(T)|_{\nabla^2 g(\xi)}\}.$$

Let

$$\alpha_k^* = \frac{\|\tilde{u}_k - u_k\|^2_{\nabla^2 h}}{|\tilde{x}_k(T) - x_k(T)|^2_{\nabla^2 g(\xi)} + \|\tilde{x}_k - x_k\|^2_{\nabla^2 f(t,\zeta)} + \|\tilde{u}_k - u_k\|^2_{\nabla^2 h}}. \tag{2.7}$$

Then $0 < \alpha_k^* < 1$[1] and $\varphi(\alpha_k^*) = 0$. It therefore follows from (2.6) that

$$c(u_k + \alpha_k^*(\tilde{u}_k - u_k)) - m \leqq (1 - \alpha_k^*)(c(u_k) - m).$$

This inequality, together with the definition of u_{k+1}, shows that

$$c(u_{k+1}) - m \leqq (1 - \alpha_k^*)(c(u_k) - m). \tag{2.8}$$

We wish now to show that there exists a constant $\alpha > 0$ such that $\alpha_k^* \geqq \alpha$ for $k = 1, 2, \cdots$. The geometric convergence of the values $c(u_k)$ to m will then follow from (2.8).

Since the set Ω is compact and the Hessians $\nabla^2 g$, $\nabla^2 f$ are continuous, we can find positive constants Λ_g and Λ_f (independent of k) such that

$$|\tilde{x}_k(T) - x_k(T)|^2_{\nabla^2 g(\xi)} \leqq \Lambda_g|\tilde{x}_k(T) - x_k(T)|_n^2 \tag{2.9}$$

and

$$\|\tilde{x}_k - x_k\|^2_{\nabla^2 f(t,\zeta)} \leqq \Lambda_f\|\tilde{x}_k - x_k\|_I^2. \tag{2.10}$$

I denotes the $n \times n$ identity matrix.

Let $\Delta x = \tilde{x}_k - x_k$ and let $\Delta u = \tilde{u}_k - u_k$. We then have

$$\Delta\dot{x} = A\Delta x + B(\nabla^2 h)^{-1/2}(\nabla^2 h)^{1/2}\Delta u,$$

from which it follows that

$$\begin{aligned} |\Delta x(t)| &\leqq \int_0^t |A(\tau)|_n|\Delta x(\tau)|_n\, dt + \int_0^T |B(\nabla^2 h)^{-1/2}|_m|(\nabla^2 h)^{1/2}\Delta u|_m\, d\tau \\ &\leqq \int_0^t |A(\tau)|_n|\Delta x(\tau)|_n\, d\tau + \|\Delta u\|_{\nabla^2 h}\|B\|_{(\nabla^2 h)^{-1}}. \end{aligned}$$

It now follows from Gronwall's inequality [18, p. 39] that

$$|\Delta x(t)|_n \leqq \|\Delta u\|_{\nabla^2 h}\|B\|_{(\nabla^2 h)^{-1}}\left(1 + \int_0^t \exp\left(\int_\tau^t |A(s)|_n\, ds\right) d\tau\right).$$

[1] If $\tilde{u}_k = u_k$, then it follows from (2.5) that $c(u_k) \leqq m$. This says that u_k is optimal.

This shows that there exists a constant r such that

$$|\Delta x(t)|_n \leqq r\|\Delta u\|_{\nabla^2 h} \tag{2.11}$$

for all $t \in [0, T]$. It follows from (2.9), (2.10) and (2.11) that

$$|\Delta x(T)|^2_{\nabla^2 g} \leqq \Lambda_g r^2 \|\Delta u\|^2_{\nabla^2 h} \tag{2.12}$$

and

$$\|\Delta x\|^2_{\nabla^2 f} \leqq \Lambda_f \int_0^T |\Delta x(t)|_n^2 \, dt \leqq \Lambda_f r^2 T \|\Delta u\|^2_{\nabla^2 h}. \tag{2.13}$$

It now follows from (2.7), (2.12) and (2.13) that

$$\alpha_k^* \geqq \alpha \equiv \frac{1}{\Lambda_g r^2 + \Lambda_f r^2 T + 1} > 0 \tag{2.14}$$

for $k = 1, 2, \cdots$. This together with (2.8) shows that the sequence $\{c(u_k)\}$ converges to m at the rate of a geometric progression.

To obtain the pointwise convergence of the u_k to u^*, let ψ^* be a function satisfying

$$\dot{\psi}^* = -A^{\perp}\psi^* + \nabla f(t, x^*), \qquad \psi(T) = -\nabla g(x^*(T)). \tag{2.15}$$

Then by Lee's lemma [14, p. 242],

$$c(u) - c(u^*) \geqq \int_0^T B^{\perp}\psi^* \cdot u^* - h(t, u^*) - \{B^{\perp}\psi^* \cdot u - h(t, u)\} \, dt \tag{2.16}$$

for any $u \in \mathscr{C}$. According to Pontryagin's maximum principle, the integrand in (2.16) is $\geqq 0$ for almost all $t \in [0, T]$. This implies that

$$(B^{\perp}\psi^* - \nabla h(t, u^*(t))) \cdot (u - u^*(t)) \leqq 0 \tag{2.17}$$

for all $u \in \Omega$ and for almost all $t \in [0, T]$. To see this, note that the sign of

$$B^{\perp}\psi^* \cdot (u^* + \alpha(u - u^*)) - h(t, u^* + \alpha(u - u^*)) - (B^{\perp}\psi^* \cdot u^* - h(t, u^*))$$

is the same as that of (2.17) for sufficiently small $\alpha \in (0, 1)$, and any $u \in \Omega$. If now in (2.16) we take $u = u_k$, we obtain, by (2.17),

$$c(u_k) - c(u^*) \geqq \tfrac{1}{2}\|\Delta u_k\|^2_{\nabla^2 h},$$

where $\Delta u_k = u_k - u^*$. This shows that the sequence $\{\|\Delta u_k\|^2_{\nabla^2 h}\}$ converges to zero at the rate of a geometric progression. Now let ψ_k be a function satisfying

$$\dot{\psi}_k = -A^{\perp}\psi_k + \nabla f(t, x_k), \qquad \psi_k(T) = -\nabla g(x_k(T)).$$

Then according to Pontryagin's maximum principle the function $\tilde{u}_k$ satisfies

$$\max_{u \in \Omega} B^{\perp}\psi_k \cdot u - h(t, u) = B^{\perp}\psi_k \cdot \tilde{u}_k(t) - h(t, \tilde{u}_k(t))$$

for almost all $t \in [0, T]$. We therefore have

$$(B^{\perp}\psi_k - \nabla h(t, \tilde{u}_k)) \cdot (u - \tilde{u}_k) \leqq 0 \tag{2.18}$$

for all $u \in \Omega$ and for almost all $t \in [0, T]$. Let us now recall that h is quadratic in u.

We can therefore write

$$h(t,u) = h(t,0) + \nabla h(t,0)\cdot u + \tfrac{1}{2}u\cdot(\nabla^2 h)u.$$

Now taking $u = \tilde{u}_k(t)$ in (2.17) we can write

$$\begin{aligned}(B^{\perp}\psi^* - \nabla h(t,0) - (\nabla^2 h)u^*)\cdot(\tilde{u}_k - u^*)\\ = ((\nabla^2 h)^{-1/2}(B^{\perp}\psi^* - \nabla h(t,0)) - (\nabla^2 h)^{1/2}u^*)\\ \cdot(\nabla^2 h)^{1/2}(\tilde{u}_k - u^*) \leqq 0.\end{aligned} \tag{2.19}$$

Similarly, from (2.18) with $u = u^*(t)$ we have

$$((\nabla^2 h)^{1/2}\tilde{u}_k - (\nabla^2 h)^{-1/2}(B^{\perp}\psi_k - \nabla h(t,0)))\cdot(\nabla^2 h)^{1/2}(\tilde{u}_k - u^*) \leqq 0. \tag{2.20}$$

Adding (2.19) and (2.20) we obtain

$$|\Delta\tilde{u}_k(t)|^2_{\nabla^2 h} \leqq |(\nabla^2 h)^{-1/2}B^{\perp}(\psi_k - \psi^*)|_m |\Delta_k\tilde{u}(t)|_{\nabla^2 h},$$

where $\Delta\tilde{u} = \tilde{u}_k - u^*$. We rewrite this as

$$|\Delta\tilde{u}_k(t)|_{\nabla^2 h} \leqq |(\nabla^2 h)^{-1/2}B^{\perp}(\psi_k(t) - \psi^*(t))|_m. \tag{2.21}$$

From (2.11) applied to Δu_k it follows that the sequence $\{|x_k(t) - x^*(t)|_n\}$ converges uniformly and geometrically to zero for each $t\in[0,T]$. And from this it follows readily that the sequence $\{|\psi_k(t) - \psi^*(t)|_n\}$ has the same property. And because of (2.21), the same can be said for the sequence $\{|\tilde{u}_k(t) - u^*(t)|_m\}$. If we observe now that

$$u_{k+1}(t) - u^*(t) = (1 - \alpha_k)(u_k(t) - u^*(t)) + \alpha_k(\tilde{u}_k(t) - u^*(t)),$$

the pointwise convergence of $\{u_k\}$ to u^* becomes clear.

Remarks. It should be observed that the only requirement on α_k for our discussion to be valid is that $c(u_{k+1}) \leqq c(u_k + \alpha_k^*(\tilde{u}_k - u_k))$, where α_k^* is given by (2.7). This suggests that the one-dimensional minimization in (iii) need not be carried out to a high degree of accuracy. It should also be noted that in the case where the functions g and f are quadratic forms in x, (2.7) determines α_k^* explicitly. And in this case we can take $\alpha_k = \alpha_k^*$. This choice of α_k is used for the example discussed in the next section.

3. Some computational results. In this section, we illustrate how the iterative procedure (i), (ii) and (iii) works for a particular example. For simplicity, we take a model which is quite familiar in the literature. It is described, for example, by Craig and Flügge-Lotz in [15].

The problem to be considered is that of optimally controlling the yaw, roll and pitch motions of a spacecraft. The small-angle variations of these motions can be described by the set of differential equations:

(3.1a) Yaw: $I_1(\ddot{\psi} - \omega\varphi) + \omega(I_3 - I_2)(\dot{\varphi} + \omega\psi) = F_1,$

(3.1b) Roll: $I_2(\ddot{\varphi} + \omega\dot{\psi}) + \omega(I_1 - I_3)(\dot{\psi} - \omega\varphi) + 3\omega^2(I_3 - I_1)\varphi = F_2,$

(3.1c) Pitch: $I_3\ddot{\theta} + 3\omega^2(I_2 - I_1)\theta = F_3.$

I_1, I_2 and I_3 are the moments of inertia about the yaw, roll and pitch axes, respectively. ω is the orbital angular frequency.

We are tacitly making certain assumptions about the spacecraft that are spelled out more clearly in [15] and [16]. For example, we are assuming that the spacecraft is traveling in a circular orbit about the earth and that we can effect twisting torques about the yaw, roll and pitch axes of the spacecraft, in an effort to stabilize it. These torques are denoted by F_1, F_2 and F_3 in (3.1). Usually they are produced by gas jets.

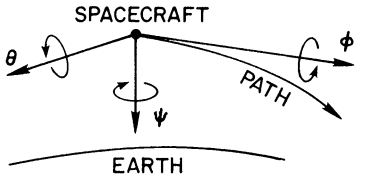


FIG. 1

The circular arrows in Fig. 1 indicate positive directions for the yaw, roll and pitch.

Note that the pitch equation (3.1c) is decoupled from the other two equations and therefore the pitch can be controlled as a separate entity. We assume this to be done and consider the harder problem of controlling the yaw and roll motions.

As in [15], we make the changes in variables:

$$\frac{F_1}{\omega^2 I_2} = u_1, \qquad \frac{F_2}{\omega^2 I_2} = u_2,$$

$$\frac{I_3 - I_2}{I_1} = a_1, \qquad \frac{I_3 - I_1}{I_2} = a_2,$$

$$x_1 = \psi, \quad x_2 = \dot{x}_1, \quad x_3 = \varphi, \quad x_4 = \dot{x}_3.$$

Now, after an appropriate scaling of time, the system (3.1a)–(3.1b) can be written as

$$\text{(3.2)} \qquad \frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -a_1 & 0 & 0 & (1 - a_1) \\ 0 & 0 & 0 & 1 \\ 0 & (a_2 - 1) & -4a_2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

The problem of optimal control is to determine an admissible control which tends to reduce the deviations x_1, x_2, x_3, x_4 to zero in a specified time while using a minimum expenditure of fuel. It is shown in [17] that the fuel expended in a time interval of length T is proportional to

$$\int_0^T \{u_1^2(t) + u_2^2(t)\}\, dt.$$

Thus a reasonable performance criterion for our system is

$$c(u) = \int_0^T \{x_1^2(t) + x_2^2(t) + x_3^2(t) + x_4^2(t) + u_1^2(t) + u_2^2(t)\}\, dt.$$

We shall take $T = 5$.

According to [15], $a_1 = a_2 = .5$ represent realistic values for the coefficients in (3.2). This assumes that the unit of time is one second. Also, a reasonable class of admissible controls is provided by the measurable functions $u = (u_1, u_2)$ on $[0, T]$ which satisfy $|u_1| \leqq .1$ and $|u_2| \leqq .1$. Figure 2 shows the system of optimal tra-

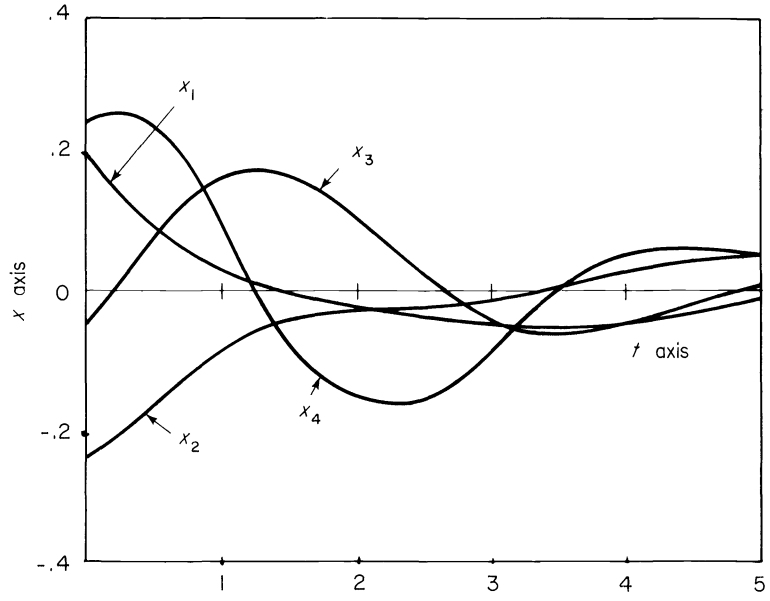


FIG. 2

jectories obtained for the starting conditions

$$x_1(0) = .20, \qquad x_2(0) = -.25,$$
$$x_3(0) = -.05, \qquad x_4(0) = .25.$$

Figure 3 shows the corresponding optimal controls. Table 1 gives the α_k computed according to (2.7) at each step of the algorithm. We started the iterative procedure with initial controls $u_1(t) = u_2(t) = 0$. Observe that convergence occurred in 21 iterations. The Frank–Wolfe algorithm required 63 iterations for convergence from the same initial conditions. One would not expect the Frank–Wolfe algorithm to converge rapidly for this example since the constraint set is not strongly convex. Here convergence is defined by the requirement that $\max_t |x_{k+1}(t) - x_k(t)| < 10^{-3}$, $t \in [0, 5]$. Calculations were done on an IBM 360/91.

We have shown how to compute a control which steers the system (3.2) to a neighborhood of the origin. Before we can claim that this is a useful control we

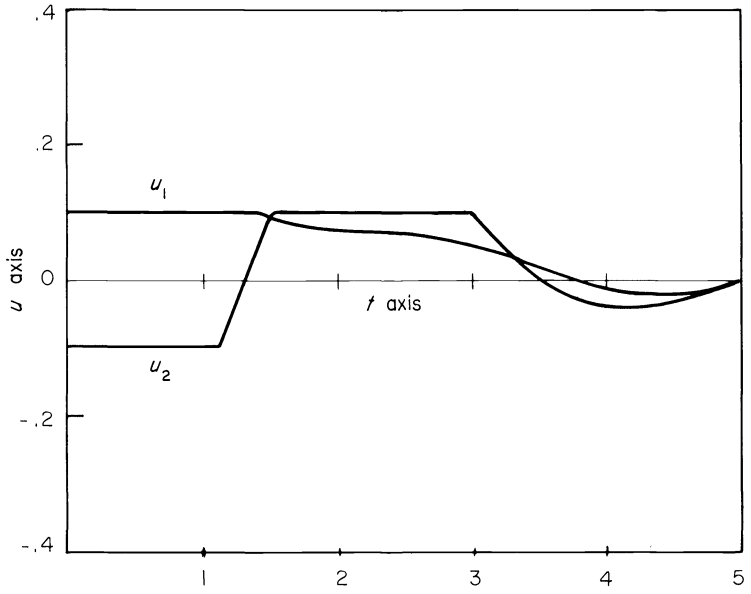


FIG. 3

k	α_k	k	α_k
1	.131345	12	.236201
2	.129714	13	.288262
3	.127972	14	.309256
4	.125988	15	.385585
5	.124503	16	.438479
6	.124881	17	.431125
7	.127727	18	.162901
8	.133701	19	.579551
9	.144472	20	.175627
10	.162005	21	.530804
11	.190451		

must verify that the system remains close to the origin after control ceases to be applied, i.e., we must verify that the homogeneous part of (3.2) is stable. But this is trivial since it is easy to show that the matrix of the homogeneous part has only purely imaginary eigenvalues.

Remark. The results of this paper have certain implications for systems with distributed parameters. This problem will be treated elsewhere.

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. 61–79.

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

[3] M. D. CANON AND C. D. CULLUM, *A tight upper bound on the rate of convergence of the Frank–Wolfe algorithm*, this Journal, 6 (1968), pp. 509–516.

[4] E. S. LEVITIN AND B. T. POLYAK, *Constrained minimization methods*, Zh. Vychisl. Mat. i Mat. Fiz., 6 (1966), no. 5, pp. 787–823 = U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), no. 5, pp. 1–50.

[5] R. O. BARR, *Computation of optimal controls by quadratic programming on convex reachable sets*, Doctoral thesis, University of Michigan, Ann Arbor, 1966.

[6] ———, *Computation of optimal controls on convex reachable sets*, Mathematical Theory of Control, Academic Press, New York, 1967, pp. 63–70.

[7] ———, *An efficient computational procedure for a generalized quadratic programming problem*, this Journal, 7 (1969), pp. 415–429.

[8] E. R. BARNES, *An extension of Gilbert's algorithm for computing optimal controls*, J. Optimization Theor. Appl., 7 (1971), pp. 420–443.

[9] M. VALADIER, *Extension d'un algorithme de Frank et Wolfe*, Rev. Française de Recherche Operationnelle, 36 (1965), pp. 251–253.

[10] A. AUSLENDER AND F. BRODEAU, *Convergence d'un algorithme de Frank et Wolfe applique a un problem de controle*, Rev. Française d'Informatique et de Recherche Operationnelle, 7 (1968), pp. 3–12.

[11] V. F. DEM'YANOV, *The solution of some optimal control problems*, this Journal, 6 (1968), pp. 59–72.

[12] L. CESARI, *Existence theorems for optimal solutions in Pontryagin and Lagrange problems*, this Journal, 3 (1966), pp. 475–498.

[13] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[14] E. B. LEE, *A sufficient condition in the theory of optimal control*, this Journal, 1 (1963), pp. 241–245.

[15] A. J. CRAIG AND I. FLÜGGE-LOTZ, *Investigation of optimal control with a minimum fuel consumption criterion for a fourth-order plant with two control inputs: synthesis of an efficient sub-optimal control*, Joint Automatic Control Conference, Stanford University, Stanford, Calif., 1964, pp. 207–221.

[16] D. B. DEBRA, *The large attitude motions and stability, due to gravity, of a satellite with passive damping in an orbit of arbitrary eccentricity about an oblate body*, SUDAER Rep. 126, Stanford University, Stanford, Calif., 1962.

[17] J. H. IRVING, *Low thrust flight: variable exhaust velocity in gravitational fields*, Space Technology, H. S. Seifert, ed., John Wiley, New York, 1969, Chap. 10.

[18] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

MAXIMUM PRINCIPLE FOR SEMICONVEX PERFORMANCE FUNCTIONALS*

ANDRZEJ P. WIERZBICKI†

Abstract. An optimal control problem is considered where the performance functional is a semiconvex function of integral functionals, i.e., a function which can be approximated locally by a convex cone. A discussion of various types of conical approximations and a variant of the basic separation lemma are presented. A maximum principle is formulated and an extension of the basic proof of the maximum principle is given. An example shows the possible applications of the variant of the maximum principle.

1. Introduction. In some practical cases an optimization problem can consist of finding a minimum of a functional of the form

$$Q = g(x_g(t_1), x(t_1)), \tag{1}$$

where g is a scalar function, $x_g(t)$ is a p-vector of "performance state" defined by

$$x_g(t) = \int_{t_0}^{t} f_g(x(\tau), u(\tau), \tau)\, d\tau, \tag{2}$$

$x(t)$ is the n-vector of state

$$x(t) = x(t_0) + \int_{t_0}^{t} f(x(\tau), u(\tau), \tau)\, d\tau \tag{3}$$

and $u(t)$ is the m-vector of control; the instant t_1 is given explicitly or implicitly, or "free", i.e., given implicitly by the condition of minimizing (1) only.

This is actually a Mayer problem. The application of the maximum principle as a necessary condition of optimality in this problem is known—provided the function g is differentiable. The following question remains: what are the minimum assumptions concerning the function g which allow formulation and proof of a maximum principle for the problem?

The question is strongly related to some problems of modern variational theory [1], [2], [3], but can also be approached in terms of more widely known formulations of the maximum principle [4], [5].

2. Conical and convex approximations. In order to establish the minimal assumptions to be fulfilled by the function g, it is useful to investigate some conical approximations of the level sets of the function. There are many types of conical approximations; for example, a variety of them are discussed in [7]. To prepare the ground for some new notions, part of the discussion must be repeated here. The discussion will be carried out in R^n space, although most of the notions and proper-

* Received by the editors March 29, 1971, and in revised form August 16, 1971.

† Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota. Now at Institute of Automatics, Department of Electronics, Warsaw University of Technology, Warsaw, Poland. This work was supported by the International Research and Exchanges Board.

ties of conical approximations can be formulated also in linear topological vector spaces.

We shall say that a cone is an *open cone* if it consists of its nonempty interior and its vertex (hence it is not precisely an open set). Recall the definition of an internal cone (see [2], [3]; in [7] another definition is used).

DEFINITION 1. The *internal cone* $\mathrm{IC}(A, x^*)$ of a set A in R^n at a point $x^* \in A$ is a nontrivial convex cone such that if a vector $r \in \mathrm{IC}(A, x^*)$, then there exist

(a) an open cone $\mathrm{OC}(r)$ such that $r \in \mathrm{OC}(r) \subset \mathrm{IC}(A, x^*)$; and

(b) a neighborhood $N(r)$ of x^* such that $\{x^* + \mathrm{OC}(r)\} \cap N(r) \subset A$.

Clearly, an internal cone is an open cone and a set A can have an internal cone only if it has nonempty interior.

Consider the class $\mathscr{A}$ of sets R^n with nonempty interior either in R^n or relative to a k-dimensional linear manifold M^k in R^n, $1 \leqq k < n$ (i.e., of sets A which contain points x such that there are neighborhoods N of x such that either $N \subset A$ or $N \cap M^k = A$). Denote by $\bar{A}$ the closure of A, by $\tilde{A}$ the interior or relative interior of A. Consider the subclass $\tilde{\mathscr{A}}$ of sets $A \in \mathscr{A}$ such that $\tilde{\bar{A}} = A$.

We shall say that a cone is a relative open cone if it consists of its nonempty relative interior and its vertex. We can define now a relative internal cone $\mathrm{RIC}(A, x^*)$ by weakening the condition (a) of Definition 1 and assuming that $\mathrm{OC}(r)$ is a relative open cone. Clearly, the internal cone can be considered to be a special, full dimensional case of the relative internal cone.

If a set $B \notin \tilde{\mathscr{A}}$, a weaker type of approximation is used. We shall say that a set is a (k-dimensional) *simplex* if it is the convex hull of the set $\{0, r_1, \cdots, r_i, \cdots, r_k\}$, where r_i are linearly independent vectors and $k \geqq 1$. The following definition of a *simplicial linearization* is equivalent to that given in [3]; with slight changes of definition, the names *conical approximation of the second kind* [6], [7] and *first order, convex approximation* [2] are also used.

DEFINITION 2. A *simplicial linearization* $\mathrm{SL}(B, x^*)$ of a set B in R^n at a point $x^* \in B$ is a convex set containing at least one nontrivial simplex and such that for any simplex $S \subset \mathrm{SL}(B, x^*)$ there is a number ε_1 and a map ζ from $\{x^*\} + S$ into R^n, possibly depending on ε_1, such that for $\varepsilon \in (0, \varepsilon_1)$ and $r \in S$:

(a) $\zeta(x^* + \varepsilon r) \in B$;

(b) $\zeta(x^* + \varepsilon r) - x^* - \varepsilon r = o(\varepsilon)$; and

(c) ζ is continuous.

The last property is particularly important.

We shall say that a cone is a *regular approximating cone* $\mathrm{RAC}(B, x^*)$ if it is a simplicial linearization (it is equivalent to the conical approximation of the second kind, whereas the conical approximation of the first kind (see [6], [7]) is equivalent to the relative internal cone). It is easy to check that every relative internal cone is a regular approximating cone with ζ being the identity map.

A converse statement is not easy to check and has played an important role in the theory of optimal control.

LEMMA 1. *If a regular approximating cone* $\mathrm{RAC}(A, x^*)$ *has nonempty interior* $\widetilde{\mathrm{RAC}}(A, x^*)$ *in* R^n, *then the interior is an internal cone* $\mathrm{IC}(A, x^*)$.

Proof. Since an equivalent statement has actually been proven in almost every known proof of the maximum principle, a shortened version of proof is presented

here. Suppose $r_1 \in \widetilde{\widetilde{\mathrm{RAC}}}(A, x^*)$; we shall construct an open cone $\mathrm{OC}(r_1)$ that fulfills the conditions (a) and (b) of Definition 1 for $\mathrm{IC}(A, x^*) = \widetilde{\widetilde{\mathrm{RAC}}}(A, x^*)$.

There is an n-dimensional simplex $S_1 \subset \widetilde{\widetilde{\mathrm{RAC}}}(A, x^*)$ such that $r_1 \in S_1$. Choose ε_1 according to Definition 2 and consider the simplex $S_2 = \varepsilon_1 S_1$ and the vector $r_2 = \varepsilon_1 r_1$. Choose a radius δ such that a ball centered at r_2 and with the radius 2δ is contained in S_2. Consider the ball $B = \{p \in R^n : \|p - r_2\| \leqq \delta\}$ and the open cone $\mathrm{OC}(r_1) = \{r \in R^n : r = \lambda p, p \in \tilde{B}, \lambda \geqq 0\}$. Clearly ((a) of Definition 1), $r_1 \in \mathrm{OC}(r_1) \subset \widetilde{\widetilde{\mathrm{RAC}}}(A, x^*)$. The property (b) of Definition 1 is equivalent to the condition that for each $p \in B$ there is an $\alpha_1 > 0$ such that $x^* + \alpha p \in A$ for $\alpha \in (0, \alpha_1)$. But for $p \in \tilde{B}$ and $\alpha \in (0, 1]$, the ball $B_{\alpha p} = \{r \in R^n : \|r - \alpha p\| \leqq \alpha\delta\} \subset S_2$. Consider the map $\xi(r) \triangleq \zeta(x^* + r)$ from $B_{\alpha p}$ into R^n, ζ being defined as in Definition 2. The map ξ has the following properties: ((a) of Definition 2) $\xi(r) \in A$ for $r \in B_{\alpha p} \subset S_2$; ((b) of Definition 2) there is an α_1 such that $\|\xi(r) - x^* - r\| < \|r - \alpha p\| \leqq \alpha\delta$ for $r \in B_{\alpha p}$, $\alpha \in (0, \alpha_1)$; ((c) of Definition 2) $\xi(r)$ is continuous. Hence for $p \in B$ and $\alpha \in (0, \alpha_1)$, from (b) and (c) of Definition 2 and the Brouwer fixed-point theorem applied to the map ξ, there is an $r \in B_{\alpha p}$ such that $\xi(r) = x^* + \alpha p$; from (a) of Definition 2, $x^* + \alpha p \in A$. This proves the property of Definition 1. This completes the proof.

It is possible to prove an analogous lemma for the relative interior of $\mathrm{RAC}(A, x^*)$ and the relative internal cone, but only under the additional assumption that the approximated set $A \in \tilde{\mathscr{A}}$, because $\mathrm{RAC}(B, x^*) \in \tilde{\mathscr{A}}$ even if $B \notin \tilde{\mathscr{A}}$.

But there are cases when even the regular approximating cone does not exist, usually due to the discontinuity of all possible maps ζ. It is, therefore, useful to introduce a new notion of a still weaker approximation, which will be called an *external cone.*

DEFINITION 3. A convex cone $\mathrm{EC}(B, x^*)$ will be an *external cone* to the set B in R^n at the point $x^* \in B$ if for each vector $r \in \mathrm{EC}(B, x^*)$ and each open cone $\mathrm{OC}(r)$ containing the vector, there is a neighborhood $N_1(r)$ of x^* such that for any neighborhood $N \subset N_1(r)$ the set $\{x^* + \mathrm{OC}(r)\} \cap B \cap N \setminus \{x^*\}$ is not empty.

LEMMA 2. *Each internal, relative internal and regular approximating cone is an external cone.*

Proof. It is obvious for the internal and relative internal cone. To prove it for the regular approximating cone, consider a vector $r \notin \mathrm{EC}(B, x^*)$. There is an open cone $\mathrm{OC}(r)$ such that $r \in \mathrm{OC}(r)$ but $\{x^* + \mathrm{OC}(r)\} \cap B \cap N = \{x^*\}$ for some neighborhood N of x^*. Therefore, there is an $\varepsilon_1 > 0$ and a $\delta > 0$ such that $\min_{z \in B} |x^* + \varepsilon r - z| > \delta\varepsilon$ for all $\varepsilon \in (0, \varepsilon_1)$. Hence, there is no mapping ζ which would fulfill the conditions (a) and (b) in Definition 2. Therefore, if $r \notin \mathrm{EC}(B, x^*)$, then $r \notin \mathrm{RAC}(B, x^*)$, and the proof is complete.

The relation

$$\mathrm{IC}(A, x^*) \Rightarrow \mathrm{RIC}(A, x^*) \Rightarrow \mathrm{RAC}(A, x^*) \Rightarrow \mathrm{EC}(A, x^*) \tag{4}$$

expressed an order in the first, "axiomatic," family of conical, convex approximations. The definitions of these approximations are not constructive and, therefore, are difficult to apply. However, there is a second, "constructive," family of conical approximations.

DEFINITION 4. A cone $\mathrm{TC}(B, x^*)$ defined by the relation

$$\mathrm{TC}(B, x^*) = \{r \in R^n: \text{there exists } \varepsilon_1 > 0 \text{ and there exists } o(\varepsilon) \text{ such that for each } \varepsilon \in (0, \varepsilon_1),\ x^* + \varepsilon r + o(\varepsilon) \in B\} \tag{5}$$

is called the *tangent cone* of the set B at the point $x \in B$.

It is possible to prove (with the help of the argument used in the proof of Lemma 2) that each convex cone contained in a tangent cone is an external cone (even the *sequential tangent cone* (see [7]), if convex, is an external cone). But a convex tangent cone is not always a regular approximating cone, because the function $o(\varepsilon)$ is not always sufficiently regular and the condition (c) in Definition 2 of the simplicial linearization is not always fulfilled.

DEFINITION 5. A cone $\mathrm{RC}(A, x^*)$ defined by the relation

$$\mathrm{RC}(A, x^*) = \{r \in R^n: \text{there exists } \varepsilon_1 > 0 \text{ such that for each } \varepsilon \in (0, \varepsilon_1),\ x^* + \varepsilon r \in A\} \tag{6}$$

is called the *radial cone* of the set $A \in \mathscr{A}$ at the point $x^* \in A$ (see [6], [7]).

If a set $B \notin \mathscr{A}$, then $\mathrm{RC}(B, x^*)$ is usually trivial (contains only $r = 0$). Obviously, $\mathrm{RC}(A, x^*) \subseteq \mathrm{TC}(A, x^*)$ and each radial cone, if convex, is an external cone. Moreover, all internal and relative internal cones are contained in the radial cone. But even if the radial cone is convex, its (relative) interior may not be a (relative) internal cone; it may not even be a regular approximating cone, as it is shown in Fig. 1 (see also [7]). Therefore, it is useful to strengthen the notion of a radial cone.

DEFINITION 6. Consider a set A in R^n with nonempty (relative) interior $\tilde{A}$ and a point $x^* \in A$. The *strong radial cone* $\mathrm{SRC}(A, x^*)$ of the set A at the point x^* is defined by

$$\mathrm{SRC}(A, x^*) = \{r \in R^n: \text{there exists } \varepsilon_1 > 0 \text{ such that for each } \varepsilon \in (0, \varepsilon_1),\ x^* + \varepsilon r \in \tilde{A}\} \cup \{0\}. \tag{7}$$

The *weak radial cone* is the closure of the strong one, $\mathrm{WRC}(A, x^*) = \overline{\mathrm{SRC}}(A, x^*)$.

Note that the strong radial cone is a (relative) open cone. Moreover,

$$\mathrm{SRC}(A, x^*) \subseteq \mathrm{RC}(A, x^*) \subseteq \mathrm{WRC}(A, x^*) \subseteq \mathrm{TC}(A, x^*), \tag{8}$$

which expresses an order in the second, "constructive," family of conical approximations. If convex, all these cones are obviously external cones.

LEMMA 3. *If the strong radial cone* $\mathrm{SRC}(A, x^*)$ *of the set* A *at the point* x^* *is convex, then it contains all (relative) internal cones of the set at this point, it is itself a (relative) internal cone, and, therefore, a regular approximating cone.*

Proof. Suppose there is a (relative) internal cone containing a vector $r \notin \mathrm{SRC}(A, x^*)$. But $r \in \mathrm{RC}(A, x^*) \subseteq \overline{\mathrm{SRC}}(A, x^*)$, because all (relative) internal cones are contained in the radial cone. Therefore, $r \in \partial\, \overline{\mathrm{SRC}}(A, x^*)$. But the (relative) internal cone contains also a (relative) open cone $\mathrm{OC}(r)$ such that $r \in \mathrm{OC}(r)$. Therefore, the (relative) internal cone would contain also vectors that do not belong to $\overline{\mathrm{SRC}}(A, x^*)$, which is impossible. Hence, each (relative) internal cone is contained in $\mathrm{SRC}(A, x^*)$. Conversely, each (relative) open, convex cone in $\mathrm{SRC}(A, x^*)$ fulfills the definition of a (relative) internal cone: any vector r

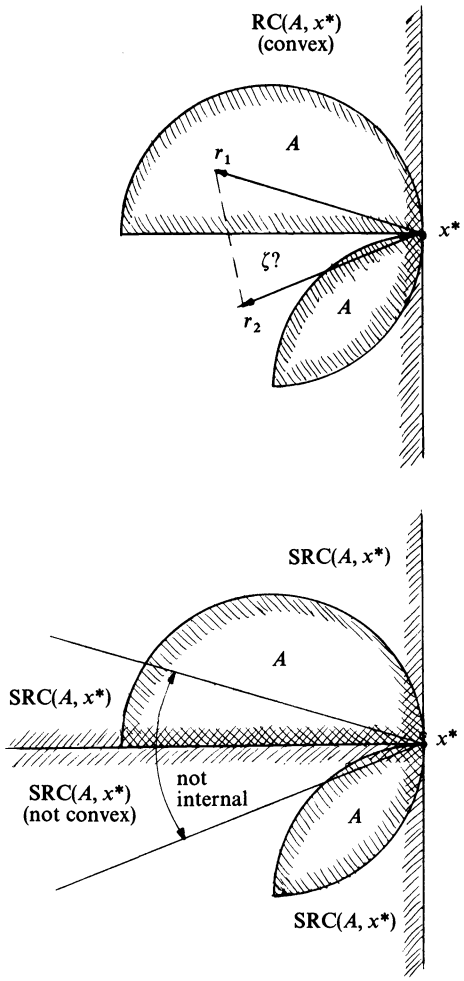


FIG. 1

in it can be contained in a (relative) open cone contained in SRC(A, x^*). But SRC(A, x^*) contains itself and is a (relative) open cone. Hence, if convex, it is a (relative) internal cone and a regular approximating cone. The proof is complete.

Some of the relations between both families of conical approximations are presented in Fig. 2.

The conical approximations of the second family are relatively easy to construct, whereas the convex cones from the first family can be better applied. Several important theorems stated with the help of these cones are presented in [2], [3], [6], [7]. Here, a slightly different result will be used.

LEMMA 4. *Consider two sets A and B in R^n such that*: (i) *A has nonempty interior $\tilde{A}$ in R^n*; (ii) *$\tilde{A} \cap B$ is empty*; (iii) *there is a point $x^* \in A \cap B$. If the set A has an internal cone* IC(A, x^*) *and the set B has an external cone* EC(B, x^*), *then the cones*

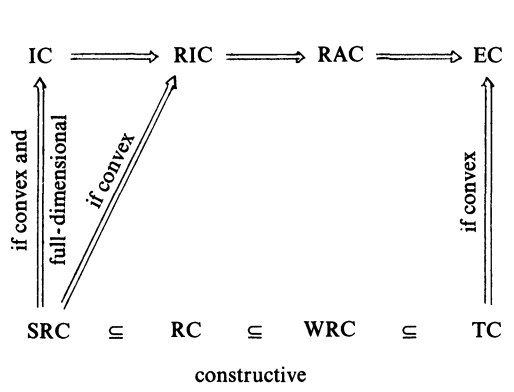


FIG. 2

$\overline{\mathrm{IC}}(A, x^*)$ *and* $\overline{\mathrm{EC}}(B, x^*)$ *can be separated, i.e., there is a hyperplane with a normal vector* η *such that* $\eta r \leqq 0$ *for all* $r \in \overline{\mathrm{EC}}(B, x^*)$ *and* $\eta l \geqq 0$ *for all* $l \in \overline{\mathrm{IC}}(A, x^*)$, *where* ηr *and* ηl *are scalar products.*

Proof. It is enough to show that $\mathrm{IC}(A, x^*) \cap \mathrm{EC}(B, x^*) = \{0\}$, because then $\tilde{\mathrm{I}}\tilde{\mathrm{C}}(A, x^*) \cap \tilde{\mathrm{E}}\tilde{\mathrm{C}}(B, x^*) = \varnothing$ which implies the separation of $\overline{\mathrm{IC}}(A, x^*)$ and $\overline{\mathrm{EC}}(B, x^*)$ by a hyperplane; see [7]. Suppose there is a nonzero vector r common to $\mathrm{IC}(A, x^*)$ and $\mathrm{EC}(B, x^*)$. Then there is an open cone $\mathrm{OC}(r)$ such that $r \in \mathrm{OC}(r) \subset \mathrm{IC}(A, x^*)$ and a neighborhood $\mathrm{NI}(r)$ of x^* such that $\{x^* + \mathrm{OC}(r)\} \cap \mathrm{NI}(r) \subset A$. Because $\mathrm{OC}(r)$ is an open cone, $\{x^* + \mathrm{OC}(r)\} \cap \mathrm{NI}(r) \setminus \{x^*\} \subset \tilde{A}$. But there is also a neighborhood $\mathrm{NE}_1(r)$ of x^* such that for any neighborhood $\mathrm{NE} \subset \mathrm{NE}_1(r)$, the set $\{x^* + \mathrm{OC}(r)\} \cap \mathrm{NE} \cap B \setminus \{x^*\}$ is not empty. Choose $\mathrm{NE} \subset \mathrm{NI}(r)$; then there is $x \neq x^*$ such that $x \in B$ and $x \in (\{x^* + \mathrm{OC}(r)\} \cap \mathrm{NE} \setminus \{x^*\}) \subset (\{x^* + \mathrm{OC}(r)\} \cap \mathrm{NI}(r) \setminus \{x^*\} \subset \tilde{A}$, which contradicts the assumption that $\tilde{A} \cap B$ is empty. Therefore, $\mathrm{IC}(A, x^*) \cap \mathrm{EC}(B, x^*) = \{0\}$, and the cones $\overline{\mathrm{IC}}(A, x^*)$ and $\overline{\mathrm{EC}}(B, x^*)$ can be separated by a hyperplane.

Observe that it would not suffice to assume that the set A has a relative internal cone, because the full dimensionality of $\mathrm{IC}(A, x^*)$ plays an important role in the proof. Although the lemma is stated in R^n, it can be generalized into a linear topological vector space. The lemma is slightly different from the results presented in [2], which imply the separation of an internal cone and a regular approximating cone.

3. Notion of semiconvexity. Instead of the function (1) of two vector arguments, it is obviously possible to consider a scalar function g of a single vector argument $x \in R^n$ (if necessary, one can introduce an augmented vector $x = (x_g(t_1), x(t_1)) \in R^{\hat{n}}, \hat{n} = p + n$). Define in R^n the set Γ_g^* of improvements of the function g with respect to a given value g^* as

$$\Gamma_g^* = \{x \in R^n : g(x) < g^*\}. \tag{9}$$

The set Γ_g^* will also be called the strong level set, in order to stress the strong inequality in its definition and to distinguish it from the usual or weak level set, with weak inequality. Obviously, Γ_g^* is open if g is continuous.

If g is convex, Γ_g^* is a convex set. If Γ_g^* is convex, the function g need not be convex. If Γ_g^* is convex for all g^*, the function g is called quasi-convex (compare, e.g., [7]). This notion of a quasi-convex function is obviously quite different and simpler than the more abstract notion of a quasi-convex family of functions (compare [1]).

Only the local properties of the set Γ_g^* are of importance when investigating the necessary conditions of optimality. Moreover, it is useful to deal with a conical approximation of this set.

Therefore, at a point $x^* \in \Gamma_g^*$, define the radial cone K_g of the set Γ_g^* as

$$K_g = \{r \in R^n: \text{ there exists } \varepsilon_1 > 0 \text{ such that for each } \varepsilon \in (0, \varepsilon_1), \\ (x^* + \varepsilon r) \in \Gamma_g^*\} \cup \{0\}, \tag{10}$$

The cone K_g will be called the *strong level radial cone*, although it is a strong radial cone only if Γ_g^* is open. But even if Γ_g^* is not open, the cone K_g has some "strong" and easy to check properties: if nonempty and convex, its relative interior forms a relative internal cone; if, in addition, full-dimensional, its interior forms an internal cone (observe that although the radial cone in Fig. 1 is convex, the corresponding strong level radial cone is not).

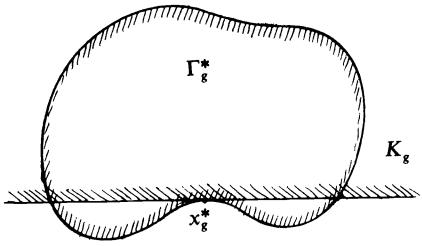
The closed cone $\overline{K}_g$ is clearly a weak radial cone and, therefore, if convex, an external cone. Obviously, the interior of $\overline{K}_g$ has different properties than K_g (see Fig. 1).

Now it is possible to define a (locally) semiconvex function g.

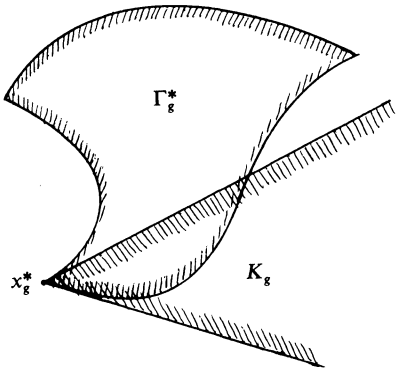
DEFINITION 7. A scalar function g defined in R^n is *weakly semiconvex* (or *shortly-semiconvex*) at a point $x^*, g(x^*) = g^*$, if there exists a nontrivial weak radial cone $\overline{K}_g$ for the level set Γ_g^* of the function and if the cone is convex (an external cone). The function g is *strongly semiconvex* at x^* if the strong level radial cone K_g is nontrivial, convex and full-dimensional in R^n (its interior is an internal cone).

Note that a function cannot be semiconvex at its minimal points, because then Γ_g^* would be empty. Furthermore, note that all differentiable functions are semiconvex (except at minimal and some stationary points), all convex and quasi-convex functions are semiconvex (except at minimal points), but there are still semiconvex functions which are neither differentiable or even locally quasi-convex at x^* (compare Fig. 3b). A function can be semiconvex even if it is not continuous. Therefore, the semiconvexity of a function at a given point is a rather weak assumption. Strong semiconvexity is a much stronger assumption, because, in addition, the cone K_g must be full-dimensional in R^n. Nevertheless, all differentiable, all locally convex and quasi-convex continuous scalar functions are strongly semiconvex (except at minimal and some stationary points).

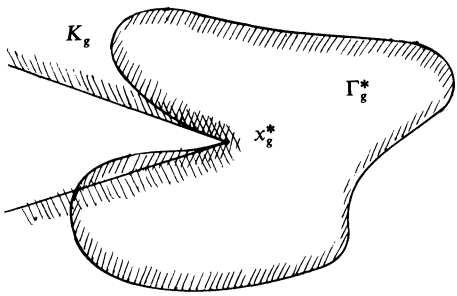
Suppose now $g = (g_1, \cdots, g_k)$ is a vector function, from R^n into R^k. Let $g(x) < g^*$ be equivalent to $g_i(x) < g_i^*$ for $i = 1, \cdots, k$. Then it is possible to define the strong level set Γ_g^* as in (9), the strong level radial cone as in (10), the weak



(a) *semiconvex*



(b) *semiconvex*



(c) *not semiconvex*

FIG. 3

radial cone and, therefore, the weakly or strongly semiconvex vector function precisely in the same way as in Definition 3.

The local convexity or quasi-convexity of all components is sufficient even for the strong semiconvexity of a vector function (provided Γ_g^* is not empty); but differentiability is not sufficient for the strong semiconvexity and additional conditions are required (they are related to the usual regularity conditions in non-linear programming).

It seems that it would be difficult to find necessary and sufficient conditions for a function to be semiconvex, i.e., to find an equivalent definition of the semiconvex function without using the notion of the strong radial level cone.

4. Notion of strong attainability. Recall the definition of the set of attainability $K(t)$ of a control process:

$$K(t) = \{x(t) \in R^n : \dot{x}(\tau) = f(x(\tau), u(\tau)),\ x(t_0) \text{ given},\ u(\tau) \in \Omega,\ \tau \in [t_0, t]\}, \tag{11}$$

where x is an absolutely continuous function of time t, u is a bounded measurable function, f and $\partial f/\partial x$ are continuous in both arguments and Ω is a compact set in R^m. Recall the definition of the tangent cone K_t to $K(t)$ at $x^*(t) \in K(t)$:

$$K_t = \mathrm{TC}(K(t), x^*(t)). \tag{12}$$

It is known that K_t has the properties of a regular approximating cone—see, e.g., [4], [5]: it is convex and the function $o(\varepsilon)$ in (5) is sufficiently regular. Therefore, it is obviously an external cone. But it does not always contain internal cones. Hence it is useful to introduce a notion of *strong attainability* in R^n.

DEFINITION 8. A point $x^*(t) \in K(t)$ is called *strongly attainable* in R^n if the tangent cone K_t at this point is full-dimensional in R^n.

Clearly, $x^*(t)$ is strongly attainable if and only if the interior of the tangent cone K_t at $x^*(t)$ is an internal cone. Observe that the *local controllability* (see [4]) implies strong attainability, but not conversely.

5. A maximum principle.

THEOREM. *Consider the control process*

$$\dot{x}(t) = f(x(t), u(t)); \qquad x(t_0) \quad \textit{given}, \tag{13}$$

where the state $x(t) \in R^n$, x is an absolutely continuous function of time $t \in [t_0, t_1]$, the control $u(t) \in \Omega \subset R^m$, Ω being compact, u is a bounded measurable function of time, and f and $\partial f/\partial x$ are continuous in R^{n+m}. Let Δ be the class of all admissible controls u that steer the initial state $x(t_0)$ to a final point $x(t_1)$ in a given final set $X_1 \subset R^n$, t_1 being free. For each u in Δ with the state response x, let the performance functional be

$$Q(u) = g(x_g(t_1)), \tag{14}$$

where g is a scalar function, $x_g(t_1) \in R^p$ is the performance state defined by

$$\dot{x}_g(t) = f_g(x(t), u(t)); \qquad x_g(t_0) \quad \textit{given}, \tag{15}$$

f_g and $\partial f_g/\partial x$ being continuous in R^{n+m}.

Let Δ_h be the subclass of Δ such that

$$h(x_h(t_1)) \leqq 0, \tag{16}$$

where h is a scalar function and $x_h(t_1) \in R^s$ is defined by

$$\dot{x}_h(t) = f_h(x(t), u(t)); \qquad x_h(t_0) \quad \textit{given}. \tag{17}$$

If u^* on $[t_0, t_1^*]$ is minimal in Δ_h and has the augmented state response $\hat{x}^* = (x_g^*, x_h^*, x^*)$ such that the function g is (weakly) semiconvex at $x_g^*(t_1^*)$ and the function h is (weakly) semiconvex at $x_h^*(t_1)^*$, then there exists a nonzero augmented adjoint response $\hat{\eta}^* = (\eta_g^*, \eta_h^*, \eta^*)$ which is a solution of the adjoint equations

$$\dot{\eta}(t) = -\eta_g(t)\frac{\partial f_g}{\partial x}(x^*(t), u^*(t)) - \eta_h(t)\frac{\partial f_n}{\partial x}(x^*(t), u^*(t)) - \eta(t)\frac{df}{\partial x}(x^*(t), u^*(t)),$$

$$\dot{\eta}_h(t) = 0; \qquad \eta_h(t) = \eta_h = \text{const.}, \tag{18}$$

$$\dot{\eta}_g(t) = 0; \qquad \eta_g(t) = \eta_g = \text{const.},$$

and such that the Hamiltonian

$$H(\hat{\eta}(t), x(t), u(t)) = \eta_g f_g(x(t), u(t)) + \eta_h f_h(x(t), u(t)) + \eta(t) f(x(t), u(t)) \tag{19}$$

is almost everywhere maximal along the optimal trajectory

$$H(\hat{\eta}^*(t), x^*(t), u^*(t)) \overset{\text{a.e.}}{=} \max_{u(t)\in\Omega} H(\hat{\eta}^*(t), x^*(t)u(t)) \tag{20}$$

and

$$\max_{u(t)\in\Omega} H(\hat{\eta}^*(t), x^*(t), u(t)) = 0 \qquad \text{on } [t_0, t_1^*]. \tag{21}$$

Suppose, additionally, that $x^*(t_1^*)$ is strongly attainable in R^{p+s+n}. If X_1 is a manifold with tangent space T_1 at $x^*(t_1^*)$, then $\eta^*(t_1^*)$ can be selected to be orthogonal to T_1 ("state transversality condition").

Furthermore, if g is differentiable at $x_g(t_1^*)$, then there exists a scalar $\lambda \leqq 0$ such that

$$\eta_g^* = \lambda\frac{\partial g}{\partial x_g}(x_g^*(t_1^*)) \tag{22}$$

("performance transversality condition"). If h is differentiable at $x_h^*(t_1^*)$, then there exists a scalar μ such that

$$\eta_h^* = \mu\frac{\partial h}{\partial x_j}(x_h^*(t_1^*)) \tag{23}$$

and

$$\mu = 0 \quad \text{if} \quad h(x_h^*(t_1^*)) < 0; \qquad \mu \leqq 0 \quad \text{if} \quad h(x_h^*(t_1^*)) = 0 \tag{24}$$

("constraint transversality condition").

Clearly, the above formulation of the maximum principle can be easily readjusted for cases when the state equations depend explicitly on time, or the final time t_1 is given, etc.

Note that the assumptions made in the first part of the theorem are much weaker than those required by Lemma 4. According to the lemma, it would be natural to assume that either the functions g and h are (mutually) strongly semiconvex at $x^*(t_1)$, or that the functions are weakly semiconvex, but the point $x^*(t_1)$ is strongly attainable. But it is possible to eliminate such assumptions when proving the first part of the theorem. First, the proof of the transversality conditions

requires additional assumptions of that kind, although one can easily imagine examples when transversality conditions are valid even if the functions f and g are only weakly semiconvex and $x^*(t_1)$ is not strongly attainable.

6. Proof. Only the necessary changes of the known proof of the maximum principle (see, e.g., [4]) are presented here.

Define the set of attainability $\hat{K}(t_1^*)$ of the augmented state $\hat{x}$ in R^{p+s+n} similarly as in (11). Define the tangent cone $\hat{K}_{t_1^*}$ of $\hat{K}(t_1^*)$ at $\hat{x}^*(t_1^*)$. To prove the existence of a nonzero augmented adjoint response resulting in the maximum principle (20) it is enough to show that $\hat{x}^*(t_1^*)$ belongs to the boundary of $\hat{K}(t_1^*)$ (see [4, Theorem 3, Chap. 4]).

If $\hat{x}^*(t_1^*)$ is not strongly attainable, then it clearly belongs to the boundary of $\hat{K}(t_1^*)$ and the maximum principle is valid—although it may constitute a trivial condition. (This can happen, for example, when one of the components of the augmented state is a linear combination of other components; but even then, in some cases, the maximum principle is nontrivial.)

If $\hat{x}^*(t_1^*)$ is strongly attainable, then the interior of the tangent cone $\hat{K}_{t_1^*}$ is an internal cone of the set $\hat{K}(t_1^*)$. Define the set of (restricted) admissible improvement $\hat{\Gamma}_{gh1}^*$ as:

$$\hat{\Gamma}_{gh1}^* = \{(x_g, x_h, x) = \hat{x} \in R^{p+s+n} : g(x_g) < g(x_g^*(t_1^*)); h(x_h) < h(x_h^*(t_1^*)); x = x^*(t_1^*)\}. \tag{25}$$

Because the functions g and h are semiconvex, there exists a weak radial cone $\hat{K}_{gh1}$ of the set $\hat{\Gamma}_{gh1}^*$ at $\hat{x}^*(t_1)$ and, being convex, it is an external cone of $\hat{\Gamma}_{gh1}^*$. The set $\hat{\Gamma}_{gh1}^*$ cannot have points in the interior of the set $\hat{K}(t_1^*)$, because that would contradict the optimality of u^*. Therefore, Lemma 4 can be applied, the cone $\hat{K}_{t_1^*}$ can be separated from $\hat{K}_{gh1}$ and cannot be the entire space R^{p+s+n}. Hence $\hat{x}^*(t_1^*)$ belongs to the boundary of $\hat{K}(t_1^*)$ and the maximum principle (20) is valid.

The set of attainability can be enlarged by time variations, $\hat{K}^{\pm}(t_1^*) = \{\hat{x} \in R^{p+n+s} : \hat{x} \in \hat{K}(t_1^* \pm \varepsilon\delta t_1); \varepsilon \in [0, \varepsilon_1]\}$. The limit cone $\hat{K}_{t_1^*}^{\pm}$ can be defined as the tangent cone of $\hat{K}^{\pm}(t_1^*)$ at $\hat{x}^*(t_1^*)$. With the help of $\hat{K}_{t_1^*}^{\pm}$, the proof of condition (21) is standard (see [4, Theorem 1, Chap. 5]).

The transversality conditions are proven under the additional assumption that $\hat{x}^*(t_1^*)$ is strongly attainable in R^{p+s+n} and, therefore, the interior of $\hat{K}_{t_1^*}$ is an internal cone of $\hat{K}^{\pm}(t_1^*)$. Define the set of (unrestricted) admissible improvement $\hat{\Gamma}_{ghx}^*$ as:

$$\hat{\Gamma}_{ghx}^* = \{(x_g, x_h, x) = \hat{x} \in R^{p+s+n} : g(x_g) < g(x_g^*(t_1^*)); \; h(x_h) < h(x_h^*(t_1^*)); \; x \in X_1\}. \tag{26}$$

The level sets of g and h have the weak radial cones $\overline{K}_g$ in R^p at $x_g^*(t_1^*)$ and $\overline{K}_h$ in R^s at $x_h^*(t_1^*)$. The set X_1 has the tangent space T_1 at $x^*(t_1^*)$. Consider the set $\hat{T}_1 = \{(r_g, r_h, r) = \hat{r} \in R^{p+s+n} : r_g \in \overline{K}_g, r_h \in \overline{K}_h, r \in T_1\}$. Clearly, $\hat{T}_1$ is a convex cone, tangent to the set $\hat{\Gamma}_{ghx}^*$ at $\hat{x}^*(t_1^*)$, and, therefore, an external cone of $\hat{\Gamma}_{ghx}^*$. Because the set $\hat{\Gamma}_{ghx}^*$ cannot have points in the interior of $\hat{K}^{\pm}(t_1^*)$, Lemma 4 can be again applied. Hence the cones $\hat{K}_{t_1^*}^{\pm}$ and $\hat{T}_1$ can be separated by a hyperplane with a normal vector $\hat{\eta}^*$ in R^{p+s+n}. Assume that $\hat{\eta}^* \hat{K}_{t_1^*}^{\pm} \leqq 0$, $\hat{\eta}^* \hat{T}_1 \geqq 0$ and take $\hat{\eta}^*$ as

the final value $\hat{\eta}^*(t_1^*)$. Because of the definition of $\hat{T}_1$, we can express $\hat{\eta}^*\hat{T}_1 = \eta_g^*\bar{K}_g + \eta_h^*\bar{K}_h + \eta^*T_1 \geqq 0$. But $0 \in \bar{K}_g$, $0 \in \bar{K}_h$, so we must have $\eta^*T_1 \geqq 0$; because T_1 is a subspace, η^* must be orthogonal to T_1.

To prove the performance transversality condition, note that if g is differentiable, then $\bar{K}_g$ is a half-space in R^p and its supporting hyperplane is uniquely determined by the normal vector $\eta_g^* = \lambda(\partial g/\partial x_g)(x_g^*(t_1^*))$, λ being a scalar. Take $\lambda \leqq 0$; then $\eta_g^*\bar{K}_g \geqq 0$ and it remains only to choose $\eta_h^*, \eta^*(t_1^*)$ such that $\hat{\eta}^*\hat{T}_1 \geqq 0$.

The same applies to the constraint transversality condition if $h(x_h^*(t_1^*)) = 0$. Clearly, if $h(x_h^*(t_1)) < 0$, then $x_h^*(t_1^*)$ is an interior point of the set $\bar{\Gamma}_h = \{x_h \in R^s: h(x_h) \leqq 0\}$, the respective cone $\bar{K}_h$ is the entire space R^s and it is necessary to assume $\eta_h^* = 0$ in order to obtain $\eta_h^*\bar{K}_h \geqq 0$.

7. An example. Consider the following problem of determining the optimal investment policy when introducing mass production and having a limited storage capacity for the ready product.

The current investment rate $u(t)$ is constrained. The total investment $x(t)$ is assumed to be equivalent to the current production. The total production is denoted by $x_g(t)$. The problem is considered on a fixed interval, $t \in [0, T]$.

The price of the ready product is a, the price connected with the investment is b. If the total amount of ready products, $x_g(t)$, is greater than a given value x_p, then an additional cost (of the storage of ready products) is taken into consideration. The price connected with the additional cost is c. The problem can be formulated as follows:

$$
\begin{aligned}
&0 \leqq u(t) \leqq 1;\\
&\dot{x}(t) = u(t); \quad x(0) = 0; \quad x(T) \text{ free};\\
&\dot{x}_g(t) = x(t); \quad x_g(0) = 0; \quad x_g(T) \text{ free};\\
&Q = g(x_g(T), x(T)) = -ax_g(T) + bx(T) + c(x_g(T) - x_p)\mathbf{1}(x_g(T) - x_p),
\end{aligned}
\tag{27}
$$

where Q is to be minimized, $\mathbf{1}(\cdot)$ is the Heaviside function, and the constants a, b, c and T are positive. In order to obtain negative values of Q (positive financial results) it is necessary to assume $aT > b$. Note that $\frac{1}{2}T^2$ is the maximal possible value of $x_g(T)$; therefore, it is assumed that $x_p < \frac{1}{2}T^2$.

The function g is not differentiable at $x_g(T) = x_p$, but it is convex. Therefore, the discussed variant of the maximum principle can be applied. The Hamiltonian

$$H = \eta_g x(t) + \eta(t)u(t) \tag{28}$$

has a unique maximum at

$$u^*(t) = \mathbf{1}(\eta(t)). \tag{29}$$

The adjoint equations are

$$\dot{\eta}(t) = \eta_g; \qquad \eta_g = \text{const.}, \tag{30}$$

which implies

$$\eta(t) = \eta(0) - \eta_g t. \tag{31}$$

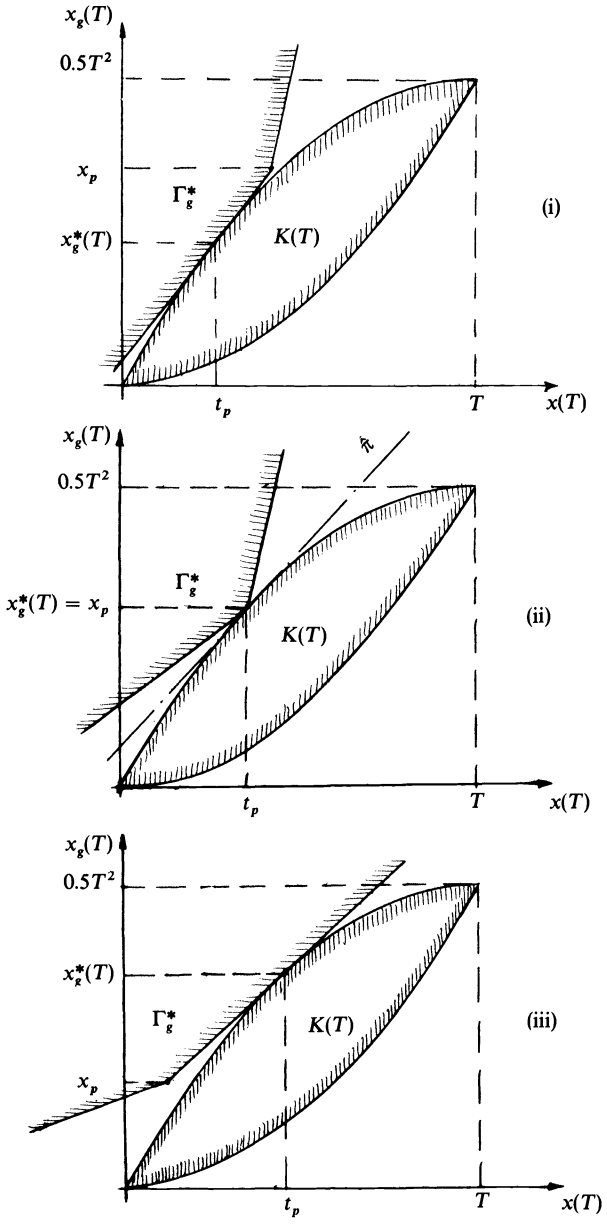


FIG. 4

Three cases of final conditions are considered—see Fig. 4:

(i) $x_g(T) < x_p$,
(ii) $x_g(T) = x_p$,
(iii) $x_g(T) > x_p$.

(i) In case (i), the transversality condition can be written as

$$\eta_g = a; \qquad \eta(T) = -b, \tag{32}$$

which implies

$$\eta(0) = -b + aT > 0. \tag{33}$$

Therefore, the optimal control is

$$u^*(t) = \begin{cases} 1, & t \in [0, t_p], \\ 0, & t \in (t_p, T], \end{cases} \tag{34}$$

where the condition $\eta(t_p) = 0$ determines the switching time t_p,

$$t_p = T - b/a. \tag{35}$$

Because

$$x(T) = t_p; \qquad x_g(T) = Tt_p - \tfrac{1}{2}t_p^2, \tag{36}$$

the case (i) corresponds to the solution of the problem if the price a of the product fulfills the inequality

$$b/T < a < b/\sqrt{T^2 - 2x_p}. \tag{37}$$

(ii) In case (ii), the formulation of the generalized maximum principle does not allow computation of $\eta_g, \eta(t)$ immediately. However, it is easy to check that the set of attainability $\hat{K}(T)$ has the form

$$\hat{K}(T) = \{x_g(T), x(T); \tfrac{1}{2}x^2(T) \leqq x_g(T) \leqq Tx(T) - \tfrac{1}{2}x^2(T)\}. \tag{38}$$

The set $\hat{K}(T)$ and the set $\hat{\Gamma}_g^*$, which is a convex cone

$$\hat{\Gamma}_g^* = \{x_g(T), x(T) : g(x_g(T), x(T)) < g^*\}, \tag{39}$$

are shown in Fig. 4 for cases (i), (ii) and (iii).

In case (ii), it is possible to define the tangent line π and the normal vector to the set $\hat{K}(T)$ at the point $(x_g^*(T), x^*(T))$, where

$$x^*(T) = t_p = T - \sqrt{T^2 - 2x_p}. \tag{40}$$

The outward normal vector to the set $\hat{K}(T)$ has the form

$$(\eta_g, \eta(T)) = -\alpha(-1, \sqrt{T^2 - 2x_p}); \qquad \alpha > 0, \tag{41}$$

which implies

$$\eta(0) = \eta(T) + \eta_g T = \alpha(T - \sqrt{T^2 - 2x_p}) > 0 \tag{42}$$

and, therefore, the optimal control has also the form (34), where the switching time t_p is now defined by (40). Furthermore, case (ii) corresponds to the optimal solution until the line π becomes a boundary of the cone $\hat{\Gamma}_g^*$. This implies the inequality

$$b/\sqrt{T^2 - 2x_p} \leqq a \leqq c + b/\sqrt{T^2 - 2x_p}. \tag{43}$$

(iii) In this case it is easy to check that the optimal control also has the form (34), where the switching time t_p is defined by

$$t_p = T - \frac{b}{a - c}. \tag{44}$$

The entire solution of the problem can be presented as in Table 1. The optimal global production can be computed from (36); its plot versus the price a is shown in Fig. 5.

TABLE 1

Case	Range of parameters	Optimal control	Comments
(0)	$a \leqq \frac{b}{T}$	$u^*(t) = 0, t \in [0, T]$	production unprofitable
(i)	$\frac{b}{T} < a < \frac{b}{\sqrt{T^2 - 2x_p}}$	$u^*(t) = \begin{cases} 1, t \in [0, t_p] \\ 0, t \in (t_p, T] \end{cases}$	$t_p = T - \frac{b}{a}$
(ii)	$\frac{b}{\sqrt{T^2 - 2x_p}} \leqq a \leqq c + \frac{b}{\sqrt{T^2 - 2x_p}}$	$u^*(t) = \begin{cases} 1, t \in [0, t_p] \\ 0, t \in (t_p, T] \end{cases}$	$t_p = T - \sqrt{T^2 - 2x_p}$
(iii)	$c + \frac{b}{\sqrt{T^2 - 2x_p}} < a$	$u^*(t) = \begin{cases} 1, t \in [0, t_p] \\ 0, t \in (t_p, T] \end{cases}$	$t_p = T - \frac{b}{a - c}$

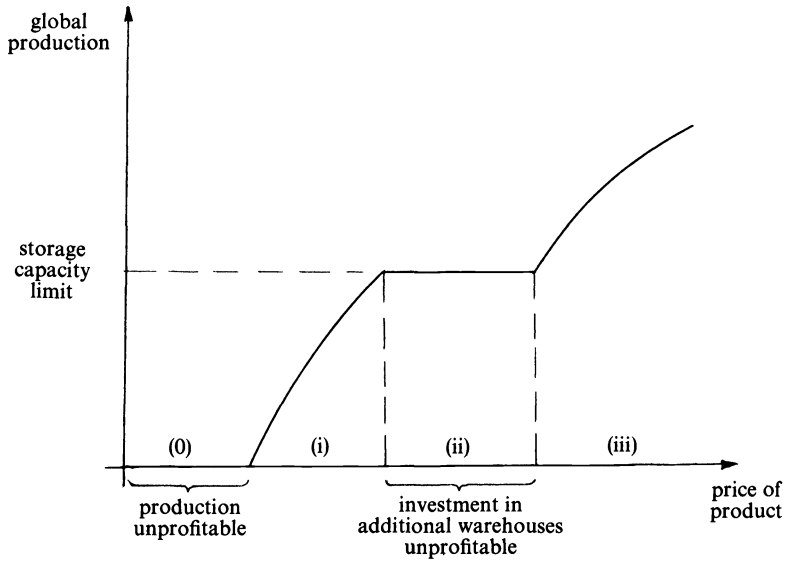


FIG. 5

8. Discussion and conclusions. If the performance functional is differentiable, then the maximum principle presented herein is only a simple and known generalization of the fundamental maximum principle. Conversely, if

$$Q(u) = g(x_g(t_1)) = x_g(t_1),$$

where $x_g(t_1)$ is a scalar, then the fundamental formulation follows immediately from the generalized one. But if the performance functional is not differentiable, the proposed generalization provides a tool to solve a broad class of problems. Besides the simple example presented in this paper, a part of this generalization was applied earlier to a problem of the optimal control of an electric arc furnace [8].

The main idea of the proposed generalization is based upon the separation of the conical approximations of the set of attainability and the set of improvement. This separation has been investigated extensively in [1], [2], [3], [6] and [7] and related papers. The discussion presented here stresses the possibility of ordering the families of conical approximations and of applying a slightly different version of the fundamental separation lemma.

Acknowledgment. The author is indebted to Professors E. Bruce Lee and Lucien W. Neustadt for their kindness in discussing the problem.

REFERENCES

[1] R. V. GAMKRELIDZE, *On some extremal problems in the theory of differential equations and applications to the theory of optimal control*, this Journal, 3 (1965), pp. 106–128.

[2] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. I: General Theory; II. Applications*, this Journal, 4 (1966), pp. 505–527; 5 (1967), pp. 90–137.

[3] ———, *A general theory of extremals*, J. Computer and System Sci., 3 (1969), pp. 57–91.

[4] E. B. LEE AND L. MARKUS, *Foundation of Optimal Control Theory*, John Wiley, New York, 1967.

[5] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[6] M. CANON, C. CULLUM AND E. POLAK, *Constrained minimization problems in finite-dimensional spaces*, this Journal, 4 (1966), pp. 528–547.

[7] ———, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.

[8] A. GOSIEWSKI AND A. WIERZBICKI, *Dynamic optimization of an electric arc furnace in a steel making plant*, Proc. 4th Congress of IFAC, Warsaw, Poland, 1968.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

THE PRIME STRUCTURE OF LINEAR DYNAMICAL SYSTEMS*

MICHAEL HEYMANN†

Abstract. In a previous paper [1] a theory, called transfer equivalence theory for constant linear dynamical systems, was presented, which led to a simple realization algorithm. In the present paper that theory is significantly sharpened by weakening the requirements for transfer equivalence, and this leads to the concepts of prime structure and prime system.

1. Introduction. Let $Z = Z(s)$ be a *proper* rational matrix; that is, each entry of Z is a quotient of real polynomials in s with the degree of the numerator lower than that of the denominator. Let (F, G, H) be a *constant linear dynamical system*; that is, F is a real square matrix, and G and H are real rectangular matrices such that the matrix product HFG is defined. The system (F, G, H) is called a *realization* of $Z(s)$, and $Z(s)$ is called the *transfer function matrix* of (F, G, H) if and only if

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

Thus, every constant linear dynamical system has a unique transfer function matrix, but every proper rational matrix has a whole class of realizations associated with it. Among the realizations of a given proper rational matrix, those systems having minimal dimension of F are called *minimal realizations.*

In an effort to investigate the intrinsic structure of linear dynamical systems and to develop an efficient theory for minimal realization of rational matrices, two concepts given in Definitions 1 and 2 were defined in [1].

DEFINITION 1. Two rational matrices Z and $\tilde{Z}$ are called *strongly equivalent* if there exist polynomial matrices $A = A(s)$ and $B = B(s)$ 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.

DEFINITION 2. Let Z and $\tilde{Z}$ be rational matrices and let $\psi = \psi(s)$ and $\tilde{\psi} = \tilde{\psi}(s)$ 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 $\psi\tilde{Z} \equiv (\psi Z)B \pmod{\psi}$.

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

In the above definition and below, two polynomial matrices $P = P(s)$ and $Q = Q(s)$ are said to be congruent modulo a given polynomial ψ, and written

* Received by the editors June 4, 1971.

† Department of Applied Mathematics, Technion—Israel Institute of Technology, Haifa, Israel.

$P \equiv Q \pmod{\psi}$, provided the corresponding entries of P and Q are congruent modulo ψ.

It is clear that both strong and weak equivalences as defined above (as well as the corresponding transfer equivalences) are equivalence relations. It is also clear that strong equivalence implies weak equivalence [1].

The definitions of strong equivalence and strong transfer equivalence were motivated by certain unsolved questions regarding the role of the numerator invariants in the Smith–McMillan canonical form [2]. This canonical form is based on the invariant factor theorem for polynomial matrices which states that given a polynomial matrix P there exist square polynomial matrices A and B with constant nonzero determinants such that $P = A\Gamma B$, where Γ is the unique polynomial matrix

$$\Gamma = \operatorname{diag}[\gamma_1, \gamma_2, \cdots, \gamma_R, 0, \cdots, 0] \tag{2}$$

with each diagonal element $\gamma_i = \gamma_i(s)$ being a monic polynomial (leading coefficient 1) which divides its successor $\gamma_{i+1}, i = 1, \cdots, R-1$.

Thus $\psi Z = A\Gamma B$, and after dividing both sides of this equation by ψ and reducing each polynomial fraction γ_i/ψ by cancellation of common factors, we have

$$Z = A\Lambda B, \tag{3}$$

where

$$\Lambda = \operatorname{diag}[\varepsilon_1/\psi_1, \cdots, \varepsilon_R/\psi_R, 0, \cdots, 0] \tag{4}$$

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

The concept of strong equivalence and strong transfer equivalence proved unsatisfactory in the study of the structure of systems in [1] since in a given class of strongly equivalent rational matrices some may be proper whereas others may not be proper and thus may not have realizations as defined above. In contrast, the concept of weak equivalence and weak transfer equivalence does not have the above mentioned handicap: For, if Z is a given rational matrix and ψ the least common denominator of the entries in Z, then if $\psi\tilde{Z} \equiv \psi Z \pmod{\psi}$ is the remainder after division by ψ of ψZ, $\tilde{Z}$ is surely proper.

With the concept of weak equivalence in mind, the central fact on which the theory in [1] hinges was stated in the following theorem.

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

$$\tilde{F} = TFT^{-1}, \quad \tilde{G} = T\sum F^j G B_j, \quad \tilde{H} = \sum A_i H F^i T^{-1} \tag{5}$$

for some nonsingular real matrix T. Then the transfer function matrix $\tilde{Z}$ of the system $(\tilde{F}, \tilde{G}, \tilde{H})$ is related to the transfer function matrix Z of (F, G, H) by

$$\psi \tilde{Z} \equiv A(\psi Z)B \,(\mathrm{mod}\, \psi), \tag{6}$$

where ψ is the minimal polynomial of F.

The above theorem led to an extensive equivalence theory and a powerful realization theory of constant linear dynamical systems. However, in an attempt to investigate the invariance properties of weak equivalence classes, difficulties were encountered (see [1, § 5a]). In particular, it was conjectured that the set $\{r, \psi_1, \cdots, \psi_r, \omega\}$ is a complete set of invariants for weak equivalence, where $\psi_1, \cdots, \psi_r$ are the polynomials which appear as denominators in the Smith–McMillan form (4), where $r = \max\{i | 1 \leqq i \leqq R,\ \psi_i \neq 1\}$, and where ω is the remainder after division by ψ_r of the polynomial product $\varepsilon_1 \cdots \varepsilon_r$ (ε_i being the numerator polynomials of the Smith–McMillan form). It was further conjectured that the canonical form associated with this set of invariants is the proper diagonal matrix

$$\Omega = \mathrm{diag}\,[1/\psi_1, \cdots, 1/\psi_{r-1}, \omega/\psi_r, 0, \cdots, 0] \tag{7}$$

which is weakly equivalent to Z (see [1, Theorem 5]).

This conjecture, which in fact can be proved to be correct, was quite surprising and unnatural. In particular, no logical explanation could be given to the appearance of the invariant ω, the physical significance of which was completely obscure. Moreover, it led to some incorrect conclusions regarding the system theoretic role of the elusive ε_i's of the Smith–McMillan form.

It will be shown below that the theory of transfer equivalence can be significantly extended by suitably modifying the definition of weak equivalence. Under this modification which involves a weakening of the requirements on the matrices A and B, the theory becomes much sharper and the difficulties encountered with the invariants are completely removed. Furthermore, new light is shed on the intrinsic structure of constant linear dynamical systems.[1]

2. Transfer equivalence. In the definition of weak equivalence as was made in [1] the matrices A and B were required to have constant nonzero determinants. Since all operations in weak equivalence classes are performed in the ring $\{\tilde{\theta}\}_\psi$, which is isomorphic to the quotient ring $R[\lambda]/R[\lambda]\psi$ (see Appendix), it is possible to weaken this requirement and demand only that A and B be unimodular in the above quotient ring, or equivalently, that A and B have determinants which are coprime with ψ. Accordingly, weak equivalence classes can be significantly extended. (Some further properties of the ring $\{\tilde{\theta}\}_\psi$ are given in the Appendix.) We, thus, make the following new definition.

---

[1] The desirability for modifying the definition of weak equivalence was independently observed by R. E. Kalman. (The point of view taken in the modified theory is also used extensively in [3] and in [4].)

DEFINITION 3. 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 $\tilde{Z}$. Then Z and $\tilde{Z}$ will be called *weakly equivalent* if and only if:

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

(ii) there exist unimodular polynomial matrices A and B over $\{\tilde{\theta}\}_{\psi}$ such that $\psi\tilde{Z} \equiv A(\psi Z)B \pmod{\psi}$.

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

Remark 1. In the above definition the term *weakly transfer equivalent* has been replaced simply by *transfer equivalent*. The reason for this is that, in view of what follows, it becomes clear that the above is the *only* natural type of transfer equivalence and that what was previously called strong transfer equivalence is quite artificial.

Since weak equivalence and transfer equivalence as defined above are equivalence relations, the equivalence theory and realization theory as developed in [1] carry over unchanged. The major improvement is in the invariant structure.

3. Prime structure of constant linear systems. In the present section we investigate the basic structural composition of constant linear dynamical systems. The basic fact is provided by the following theorem.

THEOREM 2. *Every rational matrix Z is weakly equivalent to a proper diagonal matrix Ω of the form*

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

where $\psi_1, \cdots, \psi_r$ are the nontrivial polynomials (of degree greater than zero) which appear as denominators in the Smith–McMillan form of Z. The set $\{r, \psi_1, \cdots, \psi_r\}$ forms a complete set of invariants for weak equivalence.

Proof. The proof is an immediate consequence of Theorem A2 in the Appendix.

It is well known (see, e.g., [5]) that the set of invariants $\{r, \psi_1, \cdots, \psi_r\}$ of a proper rational matrix Z is precisely the set of invariant factors of the matrix F of any minimal realization of Z. Thus, the only structural properties preserved under transfer equivalence of controllable and observable constant linear dynamical systems are the invariant factors of F and the complete controllability and complete observability properties of the system. This leads us to the following definition.

DEFINITION 4. Let $p, m, r, \psi_1, \cdots, \psi_r$ be given, where p, m and r are integers such that $p, m \geqq r$ and where $\psi_i = \psi_i(s)$ are polynomials with ψ_i dividing ψ_{i+1} for $i = 1, \cdots, r-1$. Then the $p \times m$ proper rational matrix

$$\Omega = \operatorname{diag}[1/\psi_1, \cdots, 1/\psi_r, 0, \cdots, 0] \tag{8}$$

is called the *prime rational matrix* associated with the set $\{p, m, r, \psi_1, \cdots, \psi_r\}$.

Any controllable and observable system (F, G, H) whose transfer function matrix is a prime rational matrix will be called a *prime system.*

Let Ω be a given prime rational matrix. There exists one minimal realization of Ω which is of special interest [1].

For each $i = 1, \cdots, r$, let (F_c^i, G_c^i, H_c^i) be the system

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

(9)

$$G_c^i = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix}, \qquad H_c^i = [1, 0, \cdots, 0],$$

where the a_{ij} are the coefficients of ψ_i (i.e., $\psi_i = \sum_{j=0}^{n_i} a_{ij}s^j$ where $a_{i,n_i} = 1$). The minimal realization (F_c, G_c, H_c) of Ω is a direct sum of the systems (F_c^i, G_c^i, H_c^i) augmented by zeros to give the appropriate dimensions p and m as follows:

$$F_c = \begin{bmatrix} F_c^1 & & & \\ & F_c^2 & & \\ & & \ddots & \\ & & & F_c^r \end{bmatrix}, \qquad G_c = \left[\begin{array}{cccc|c} G_c^1 & & & & \\ & G_c^2 & & & 0 \\ & & \ddots & & \\ & & & G_c^r & \end{array}\right],$$

(10)

$$H_c = \left[\begin{array}{cccc} H_c^1 & & & \\ & H_c^2 & & \\ & & \ddots & \\ & & & H_c^r \\ \hline & & 0 & \end{array}\right].$$

The realization (F_c, G_c, H_c) will be called the *canonical prime system* associated with Ω. We thus have the following theorem.

THEOREM 3 (Prime structure theorem). *Every controllable and observable constant linear dynamical system (F, G, H) is transfer equivalent to a unique canonical prime system.*

In view of the above discussion, and in particular the prime structure theorem, it is clear that every constant linear system may be viewed as a cascade composition of an "input structure modifier," a prime system, and an "output structure modifier." (In discrete systems these structure modifiers may be regarded as "coder" and "decoder.")

It is of interest to inspect this concept in terms of a modified definition of externally equivalent systems [1].

DEFINITION 5. Two constant linear dynamical systems (F, G, H) and $(\tilde{F}, \tilde{G}, \tilde{H})$ are called *externally equivalent* if

$$\tilde{F} = TFT^{-1}, \quad \tilde{G} = T\sum F^j G B_j, \quad \tilde{H} = \sum A_i H F^i T^{-1},$$

where $A = \sum A_i s^i$ and $B = \sum B_j s^j$ are square unimodular matrices over $\{\tilde{0}\}_\psi$ ($\psi = \psi(s)$ being the minimal polynomial of F) and where T is some nonsingular real matrix.

In view of the Cayley–Hamilton theorem the relationships for external equivalence can be expressed as

$$\tilde{G} = T[G, FG, \cdots, F^{n-1}G] \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_{n-1} \end{bmatrix},$$

$$\tilde{H} = [A_0, A_1, \cdots, A_{n-1}] \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix} T^{-1}, \tag{11}$$

where n is the degree of ψ. Thus the input and output structure modifications amount to certain matrix operations on the controllability and observability matrices of (F, G, H).

Specifically, let (F_c, G_c, H_c) be a canonical prime system. Let $F = TF_cT^{-1}$ for some nonsingular matrix T, and let G and H be any matrices of the same dimensions as G_c and H_c respectively such that (F, G, H) is controllable and observable. Then (F, G, H) is externally equivalent to (F_c, G_c, H_c), and in particular H and G are related to H_c and G_c by (11) for appropriate unimodular matrices $A = \sum_{i=0}^{n-1} A_i s^i$ and $B = \sum_{j=0}^{n-1} B_j s^j$.

Remark 2. Since external equivalence hinges on transfer equivalence, it is clear that not all matrices B and A that satisfy (11) are *automatically* unimodular. To illustrate this fact let

$$F = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad G = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad H = I,$$

and let $\tilde{F}$ and $\tilde{G}$ be given by

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

Then

$$[G, FG] = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}, \qquad [\tilde{G}, \tilde{F}\tilde{G}] = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

and both systems are externally equivalent. In fact,

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

so that

$$B = \begin{bmatrix} s+1 & 1 \\ 1 & s \end{bmatrix}$$

and $\det(B) = s^2 + s - 1$ which is coprime with $\psi(s) = s^3(s-1)$.

However, there also exists the relation

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

so that

$$\hat{B} = \begin{bmatrix} s+1 & 1 \\ 1 & 1 \end{bmatrix}$$

and $\det(\hat{B}) = s$ which divides $\psi(s)$.

4. Conclusion. The present paper provides sharp insight into the intrinsic structure of constant linear systems and brings into focus the concept of a prime system. Furthermore, in the modified theory of transfer equivalence the difficulties with the invariants are completely resolved.

At present, the main application of transfer equivalence theory is the realization theory that resulted from it. It is likely, however, that the theory may have applications in other theoretical and practical problems and this is currently being investigated.

1. The theory of inverse systems (see, e.g., [6] and [7]) can be related to transfer equivalence theory and, in particular, to the prime structure.

2. It is of interest to know when a system can be regarded as completely decoupled on either the input or the output side. In this connection one would

wish to be able to characterize systems (in terms of the system equations) which satisfy

$$\psi Z \equiv (\psi\Omega)B \,(\mathrm{mod}\, \psi)$$

or which satisfy

$$\psi Z \equiv A(\psi\Omega)(\mathrm{mod}\, \psi).$$

3. In [8] a unique canonical form for constant linear systems was derived in which all redundancies in parameters were removed and the system matrices contained a minimal number of parameters (minimally parametrized systems). Clearly, a canonical prime system is minimally parametrized. Can the matrices A and B be characterized which directly yield (by external equivalence) minimally parametrized systems?

In conclusion, it is worth remarking that the theory of transfer equivalence failed in finding any system theoretic significance in the invariants ε_i of the Smith–McMillan form. There seems to be little evidence at this stage, if any, that any special significance (in the system theoretic sense) of the ε_i's should be further expected.

Appendix. In this Appendix we review certain properties of matrices over the quotient ring $R[\lambda]/R[\lambda]\psi$, where $R[\lambda]$ is the polynomial ring in one variable over R, the field of real numbers, and where $\psi = \psi(\lambda)$ is a fixed polynomial in $R[\lambda]$. The ring $R[\lambda]/R[\lambda]\psi$ is generally not a principal ideal domain and the main purpose of this Appendix is to extend the invariant factor theorem to matrices over this ring. Most of the facts reviewed in this Appendix can be found in the literature in terms of principal ideal domains and torsion modules over principal ideal domains (see, e.g., [9]).

In the quotient ring $R[\lambda]/R[\lambda]\psi$ the elements are the residue classes $\theta + R[\lambda]\psi$ for $\theta \in R[\lambda]$. In each residue class there is a unique element $\tilde{\theta}$ of least degree so that the set of polynomials $\{\tilde{\theta}\}_\psi$ is isomorphic to the set of residue classes $\{\theta + R[\lambda]\psi\}$. $\tilde{\theta}$ is the unique representative of the residue class $\theta + R[\lambda]\psi$ taken as the remainder after division by ψ of any element in the residue class. The ring $R[\lambda]/R[\lambda]\psi$ is isomorphic (as a ring) to $\{\tilde{\theta}\}_\psi$, where in $\{\tilde{\theta}\}_\psi$ addition is the same as for polynomials and multiplication is defined as $\tilde{\theta}_1 \cdot \tilde{\theta}_2 = \widetilde{\theta_1 \cdot \theta_2}$.

In the ring $\{\tilde{\theta}\}_\psi$ every element is either a unit or a zero divisor. In fact, every polynomial $\tilde{\theta}$ which is coprime with ψ is a unit, since there exist polynomials α and β such that

$$\alpha\tilde{\theta} + \beta\psi = 1 \Rightarrow \tilde{\theta}^{-1} \equiv \alpha \,(\mathrm{mod}\, \psi).$$

Similarly, every element $\tilde{\theta}$ which has a nontrivial common factor α with ψ is a divisor of zero since

$$\tilde{\theta} \cdot \left(\frac{\psi}{\alpha}\right)^{\sim} = \left(\frac{\tilde{\theta}}{\alpha}\right) \cdot \psi \equiv 0 \,(\mathrm{mod}\, \psi).$$

We also have the following theorem.

THEOREM A1. *Let $\tilde{\theta}$ be any polynomial in the ring $\{\tilde{\theta}\}_{\psi}$, and let $\alpha = (\tilde{\theta}, \psi)$ denote the greatest common factor of $\tilde{\theta}$ and ψ. Then there is a unit $u \in \{\tilde{\theta}\}_{\psi}$ such that $u \cdot \tilde{\theta} \equiv \alpha \pmod{\psi}$.*

Proof. Let $\tilde{\theta} = \alpha\tilde{\theta}'$, and let $\psi = \alpha\psi'$, where $\tilde{\theta}'$ and ψ' are coprime. Then there are polynomials γ and δ (with $\gamma \in \{\tilde{\theta}\}_{\psi}$) such that $\gamma\tilde{\theta}' + \delta\psi' = 1$. If γ is coprime with ψ, set $u = \gamma$ and we have $u\tilde{\theta} = u\alpha\tilde{\theta}' \equiv \alpha \pmod{\psi}$.

If γ is not coprime with ψ, let $\hat{\psi}$ be the greatest factor of ψ which is coprime with γ. Then ψ' divides $\hat{\psi}$, and the polynomial $\gamma + \hat{\psi}$ is a unit in $\{\tilde{\theta}\}_{\psi}$, and we have

$$u\tilde{\theta} = (\gamma + \hat{\psi})\alpha\tilde{\theta}' = \alpha\gamma\tilde{\theta}' + \psi\left(\frac{\hat{\psi}}{\psi'}\right)\tilde{\theta}' \equiv \alpha \pmod{\psi}.$$

This completes the proof.

A square matrix A over $\{\tilde{\theta}\}_{\psi}$ is called unimodular if there exists a matrix B over $\{\tilde{\theta}\}_{\psi}$ such that

$$A \cdot B \equiv B \cdot A \equiv I \pmod{\psi}.$$

We shall denote B by A^{-1} and call it the inverse of A (in the ring $\{\tilde{\theta}\}_{\psi}$).

A necessary and sufficient condition for A to be unimodular is that its determinant be a unit in $\{\tilde{\theta}\}_{\psi}$, i.e., that its determinant be coprime with ψ.

It is elementary to show that every unimodular matrix A over $\{\tilde{\theta}\}_{\psi}$ can be obtained by a sequence of elementary operations of the following types on the unit matrix:

(i) interchange two rows or columns;
(ii) multiply a row or a column by a unit in $\{\tilde{\theta}\}_{\psi}$ and reduce modulo ψ;
(iii) add a polynomial multiple of one row or column to another and reduce modulo ψ.

THEOREM A2 (Invariant factor theorem for matrices over $\{\tilde{\theta}\}_{\psi}$). *Let Q be a $p \times m$ matrix over $\{\tilde{\theta}\}_{\psi}$. Then Q has the representation*

$$Q \equiv A \,\Pi\, B \pmod{\psi}, \tag{A.1}$$

where A and B are, respectively, $p \times p$ and $m \times m$ unimodular matrices over $\{\tilde{\theta}\}_{\psi}$, and where Π is the $p \times m$ matrix

$$\Pi = \operatorname{diag}(\pi_1, \pi_2, \cdots, \pi_r, 0, \cdots, 0). \tag{A.2}$$

In (A.2) *the π_i are monic factors of ψ which are uniquely determined by Q, and each π_i divides π_{i+1} for $i = 1, 2, \cdots, r-1$.*

Proof. We shall only outline a constructive proof for the existence of the representation (A.1) which is of direct interest for construction of the canonical form for transfer equivalence. The proof of uniqueness of the matrix Π will be omitted.

Let

$$Q = \hat{A}\Gamma\hat{B}$$

be the Smith canonical representation of Q, where $\hat{A}$ and $\hat{B}$ are polynomial matrices with (real) constant nonzero determinants and where

$$\Gamma = \operatorname{diag}(\gamma_1, \gamma_2, \cdots, \gamma_R, 0, \cdots, 0)$$

is the Smith canonical form of Q.

Let $\pi_i = (\gamma_i, \psi)$ denote the monic greatest common factor of γ_i and ψ, and let r be the largest integer for which $\deg \pi_i < \deg \psi$. Clearly π_i divides π_{i+1} for $i = 1, \cdots, R-1$ since this divisibility holds for the γ_i. Furthermore $\pi_i \equiv 0 \pmod{\psi}$ for $i > r$. By Theorem A1 let $u_i \in \{\tilde{\theta}\}_\psi$ be units such that $u_i \cdot \gamma_i \equiv \pi_i \pmod{\psi}$ for $i \leqq r$, and let U be the square unimodular matrix

$$U = \operatorname{diag}(u_1, u_2, \cdots, u_r, 1, \cdots, 1).$$

Then

$$U\Gamma \equiv \Pi \pmod{\psi},$$

and therefore,

$$Q \equiv (\hat{A}U^{-1})\,\Pi\,\hat{B} \equiv (\hat{A}U^{-1})_\psi\,\Pi\,(\hat{B})_\psi \pmod{\psi},$$

where $(\hat{A}U^{-1})_\psi$ and $(\hat{B})_\psi$ are the remainders of $\hat{A}U^{-1}$ and $\hat{B}$ after division by ψ.

Set $A = (\hat{A}U^{-1})_\psi$ and $B = (\hat{B})_\psi$.

REFERENCES

[1] M. Heymann and J. A. Thorpe, *Transfer equivalence of linear dynamical systems*, this Journal, 8 (1970), pp. 19–40.

[2] B. McMillan, *Introduction to formal realizability theory*, Bell System Tech. J., 31 (1952), pp. 541–600.

[3] R. E. Kalman, *Lectures on controllability and observability*, Lecture notes delivered at Centro Internazionale Matematico Estivo (C.I.M.E.), 1968.

[4] R. E. Kalman, P. L. Falb and M. A. Arbib, *Topics in Mathematical Systems Theory*, McGraw-Hill, New York, 1969.

[5] R. E. Kalman, *Irreducible realizations and the degree of a rational matrix*, SIAM J. Appl. Math., 13 (1965), pp. 520–544.

[6] M. K. Sain and J. L. Massey, *Invertibility of linear time-invariant dynamical systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 141–149.

[7] L. A. Silverman, *Inversion of Multivariable Linear Systems*, Ibid., AC-14 (1969), pp. 270–276.

[8] M. Heymann, *A unique canonical form for multivariable linear systems*, Internat. J. Control, 12 (1970), pp. 913–927.

[9] S. Lang, *Algebra*, Addison-Wesley, Reading, Mass., 1965.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

THE "BANG-BANG" PROBLEM FOR CERTAIN CONTROL SYSTEMS IN $GL(n, R)$*

HECTOR J. SUSSMANN†

Abstract. We discuss the linear control problem $\dot{X}(t) = [A_0(t) + \sum_{i=1}^{m} u_i(t)A_i(t)]X(t)$, where $A_0, \cdots, A_m$ are $n \times n$ matrix-valued functions of time, and where $X(t) \in GL(n, R)$. We show that the set attainable from any element $M \in GL(n, R)$ at time t by "bang-bang" controls is closed, provided the following very strong assumption is satisfied: for all i, j and for all t', t'' such that $0 \leqq t' \leqq t$, $0 \leqq t'' \leqq t$, the matrices $A_i(t')$ and $A_j(t'')$ commute. We also show, by means of counterexamples, that these assumptions cannot be weakened.

1. Introduction. Recently, interest has arisen in the study of the linear control problem in manifolds and Lie groups (Brockett [1], Haynes and Hermes [4], Kučera [5] and [6]). Such a control problem is of the form

$$\dot{x}(t) = X_0(x(t)) + \sum_{i=1}^{m} u_i(t)X_i(x(t)),$$

where $\dot{x}(t)$ denotes the tangent vector to the curve $\tau \to x(\tau)$ at $\tau = t$, and where $X_0, \cdots, X_m$ are vector fields.

A particularly important case is that in which the manifold is a Lie group, and the vector fields are translation-invariant. When the Lie group is $GL(n, R)$, the problem takes the simple form

$$\dot{X}(t) = \left(A_0 + \sum_{i=1}^{m} u_i(t)A_i\right) X(t),$$

where $A_0, \cdots, A_m$ are constant matrices.

The purpose of this article is to indicate what hopes there are of building a reasonable "bang-bang" theory for this problem, and for the more general one in which the matrices A_i are time-dependent. The important issue is, as usual, to determine whether the attainable set at time T is closed. It might seem likely that the tool to be used should be some generalization of the well-known theorem of Lyapunov, which has proved so fruitful for similar problems (Lyapunov [7], Halkin [2] and [3]). However, as we shall show, not much is to be expected in this direction. We shall prove that, under certain *very restrictive* conditions, closedness of the attainable set follows by a straightforward application of Lyapunov's theorem. We shall also show that, if these conditions are weakened, it is possible to give examples of control systems for which the attainable set fails to be closed.

The new aspect that plays a fundamental role is that of the *commutativity* of the matrices A_i. The very restrictive assumptions to which we have referred are the following:

(a) The condition that[1] $[A_i(t), A_j(t')]$ should vanish for all i, j, t, t'. This will guarantee that the set attainable at time T by "bang-bang" controls is closed,

* Received by the editors February 2, 1971, and in revised form June 14, 1971.

† Department of Mathematics, University of Chicago, Chicago, Illinois 60637. This work was performed while the author was at the Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts. It was supported by the U.S. Office of Naval Research under the Joint Services Electronics Program by Contract N00014-67-A-0298-0006.

[1] We are using here the standard notation $[M, N] = MN - NM$. Thus "$[M, N] = 0$" is another way of saying that the matrices M and N commute.

provided we define a "bang-bang" control as a *measurable* function with values in the set $\{-1, 1\}$.

(b) The additional condition that the functions $A_i(t)$ should be piecewise analytic. If this is true, we shall be able to get closedness even if we restrict the class of "bang-bang" controls to *piecewise constant* functions with values in $\{-1, 1\}$. Of course, this covers the time-independent case in particular.

The main point of this paper is that these conditions *cannot* be weakened. This will be shown by giving three examples of nonclosed attainable sets. These examples cover, in our opinion, the simplest possible conceivable departures from the commutativity condition. Thus, our results constitute a rather final answer to the closedness problem.

Our results also apply to systems in which X is a *column vector in* R^n, rather than an $n \times n$ matrix (cf. Remark 1 of § 4).

2. Notations and preliminary lemmas. We shall consider the control problem

$$\dot{X}(t) = \left(A_0(t) + \sum_{i=1}^{m} u_i(t)A_i(t)\right)X(t), \tag{1}$$

where $X(t)$ belongs to $GL(n, R)$ (the set of all nonsingular real $n \times n$ matrices). The functions $A_i(t)$ are supposed to be bounded and measurable, with values in the set $M(n, R)$ of all $n \times n$ real matrices. For $A \in M(n, R)$, define the norm of A (denoted by $\|A\|$) as the supremum of $\|Ax\|$, where x ranges through all the vectors in R^n such that $\|x\| = 1$, and where, for $x \in R^n$, $\|x\|$ denotes the Euclidean norm $(\sum_{i=1}^{n} x_i^2)^{1/2}$.

We shall denote by $U(T)$, for each $T > 0$, the set of all measurable[2] functions defined in the closed interval $[0, T]$ with values in the cube $\{(u_1, \cdots, u_m): -1 \leqq u_i \leqq 1, i = 1, \cdots, m\}$. We shall denote by $UB(T)$ the subset of $U(T)$ whose elements are the "bang-bang" functions, i.e., the measurable functions $(u_1(t), \cdots, u_m(t))$ such that $u_i(t) = 1$ or $u_i(t) = -1$ for all $i = 1, \cdots, m$ and all $0 \leqq t \leqq T$. Finally, the set of all $u \in UB(T)$ that are piecewise constant will be denoted by $UBP(T)$.

It is clear that $U(T)$ is a bounded and weakly closed subset of $L_2[0, T]$, so that $U(T)$ is weakly compact. We also have the following lemma:

LEMMA 1. *$UBP(T)$ is weakly dense in $U(T)$.*

Proof. It is clearly sufficient to assume $m = 1$. Since every function in $U(T)$ can be approximated in the L_2-norm by piecewise constant functions, it follows that it will be sufficient to show that every constant function is a weak limit of elements of $UBP(T)$.

Let $u(t) \equiv r \leqq 1$, for $0 \leqq t \leqq T$. We can assume $r \geqq 0$. For each interval $I = [a, b]$, let the function f_I be defined as follows:

$$f_I(t) = \begin{cases} -1 & \text{for} \quad a \leqq t \leqq a + \frac{1}{2}(1 - r)(b - a), \\ 1 & \text{for} \quad a + \frac{1}{2}(1 - r)(b - a) < t \leqq b. \end{cases}$$

Then, clearly, $\int_a^b f_I(t)\,dt = r(b - a)$. Now define u_k (for $k = 1, 2, \cdots$) by partitioning the interval $[0, T]$ into k intervals $I_{k1}, \cdots, I_{kk}$ of length Tk^{-1}, and letting $u_k(t) = f_{I_{ki}}(t)$ for each $t \in I_{ki}$, $i = 1, \cdots, k$. It is now obvious that the functions u_k

[2] We follow the standard convention of identifying functions that are equal almost everywhere.

belong to $UBP(T)$ and that their weak limit is u. The proof of our lemma is thus complete.

Let $u \in U(T)$. Let $X(u, \cdot)$ be the solution of (1) which satisfies the initial condition $X(0) = I$ ($I = n \times n$ identity matrix). The set of all matrices $X(u, T)$, for $u \in U(T)$, is the *attainable set at time* T, and we shall denote it by $S(T)$. If we restrict ourselves to functions $u \in UB(T)$ (resp. $u \in UBP(T)$), we can similarly define the sets $SB(T)$(resp. $SBP(T)$). The union of the sets $S(t)$ for all $0 \leqq t \leqq T$ will be denoted by $S'(T)$. In a similar way, we define the sets $SB'(T), SBP'(T)$.

It is clear that no loss of generality is involved in limiting ourselves to the study of the sets attainable from the identity. Indeed, the set of matrices attainable from any other $M \in GL(n, R)$ is just the set of all products $X(u, t)M$.

LEMMA 2. *Let the functions u_k converge weakly to u. Then $\{X(u_k, t)\}$ converges uniformly to $X(u, t)$ for $0 \leqq t \leqq T$.*

Proof. For each $v \in U(T)$, we have

$$X(v, t) = I + \int_0^t \left[A_0(\tau) + \sum_{i=1}^{m} v_i(\tau) A_i(\tau) \right] X(v, \tau)\, d\tau. \tag{2}$$

Since the functions A_i are bounded, and $|v_i(\tau)| \leqq 1$, there is a constant $C > 0$ such that

$$\|X(v, t)\| \leqq 1 + C \int_0^t \|X(v, \tau)\|\, d\tau$$

for all $v \in U(T)$, and all $0 \leqq t \leqq T$.

It follows by a well-known argument that

$$\|X(v, t)\| \leqq \exp(Ct) \quad \text{for all } v, t.$$

In particular, we see that the functions $X(v, \cdot)(v \in U(T))$ are uniformly bounded. Equation (1) then implies that the derivatives of these functions are also uniformly bounded.

To show that $X(u_k, \cdot)$ converges uniformly to $X(u, \cdot)$, it is sufficient to show that every subsequence has a subsequence that converges uniformly to $X(u, \cdot)$. By the previous paragraph and the Ascoli–Arzelà theorem, every subsequence has a subsequence that converges uniformly to some function. Thus, our lemma will be proved if we show that, if $\{v_k\}$ converges weakly to v, and if $X(v_k, \cdot)$ converges uniformly to $X(\cdot)$, then $X(\cdot) \equiv X(v, \cdot)$.

Equation (2) implies that

$$X(v_k, t) = I + \int_0^t \left[A_0(\tau) + \sum_{i=1}^{m} (v_k)_i(\tau) A_i(\tau) \right] [X(v_k, \tau) - X(\tau)]\, d\tau + \int_0^t \left[A_0(\tau) + \sum_{i=1}^{m} (v_k)_i(\tau) A_i(\tau) \right] X(\tau)\, d\tau.$$

Using the weak convergence of v_k to v, and the uniform convergence of $X(v_k, \cdot)$ to $X(\cdot)$, it follows that

$$X(t) = I + \int_0^t \left[A_0(\tau) + \sum_{i=1}^{m} v_i(\tau) A_i(\tau) \right] X(\tau)\, d\tau.$$

Then, $X(t) \equiv X(v, t)$, and our lemma is proved.

COROLLARY 1. *The mapping* $u \to X(u, \cdot)$ *is continuous from* $U(T)$ *with the weak topology into the space of continuous* $M(n, R)$*-valued functions in* $[0, T]$ *with the uniform topology.*

COROLLARY 2. *The sets* $S(T)$, $S'(T)$ *are compact.*

COROLLARY 3. *The sets* $SBP(T)$, $SBP'(T)$ *are dense in* $S(T)$, $S'(T)$, *respectively.*

Proof. Corollary 1 is a restatement of Lemma 2. Corollary 2 follows from Corollary 1 and the fact that $U(T)$ is weakly compact. Finally, Corollary 3 follows from Lemma 1 and Corollary 1.

3. Closedness of the "bang-bang"-attainable set. It is clear from the preceding section that closedness of the attainable set $SB(T)$ (resp. $SBP(T)$, $SB'(T)$, $SBP'(T)$) is equivalent to the identity $S(T) = SB(T)$ (resp. $S(T) = SBP(T)$, $S'(T) = SB'(T)$, $S'(T) = SBP'(T)$).

The following theorem is a positive result in this direction and, as we shall prove in the next section, it is the best possible result of that type.

THEOREM 1. *If all the brackets* $[A_i(t), A_j(t')]$ *vanish (for all* i, j, t, t'*), then* $SB(T)$ *and* $SB'(T)$ *are closed. If, in addition, the functions* A_i *are piecewise analytic, then the sets* $SBP(T)$ *and* $SBP'(T)$ *are also closed.*

Proof. If our assumption about the brackets holds, then the solution of (1) is given by

$$X(u, t) = \exp\left(\int_0^t A_0(\tau)\, d\tau\right) \cdot \prod_{i=1}^{m} \exp\left(\int_0^t A_i(\tau) u_i(\tau)\, d\tau\right).$$

To verify this, notice that: (i) the derivative of $\exp(F(t))$ is $F'(t)\exp(F(t))$, if F is a matrix-valued function *such that* $F(t_1)$ *and* $F(t_2)$ *commute* for all t_1, t_2, and that: (ii) $\exp(M + N) = \exp(M) \cdot \exp(N)$ if M and N are two *commuting* matrices (these two facts are proved, using the power series expansion for the exponential, in exactly the same way as for the scalar case; the commutativity makes it possible to "rearrange" factors). From (i) it follows easily that $X(u, t) = \exp(\int^t (A_0 + \sum u_i A_i))$. The desired expression then results from (ii).

It follows from Lyapunov's theorem on the range of a vector-valued measure (Lyapunov [7], Halkin [2]), that the set of matrices $\int_0^t A_i(\tau)u(\tau)\, d\tau$, where u ranges over the set of all measurable functions with values in $\{-1, 1\}$, is compact for each i. Thus, the first part of our statement is clear. The second part follows in a similar way: according to a theorem of Halkin (see [2] and [3]), the set of values $\int_0^t f(\tau)u(\tau)\, d\tau$, where u ranges over all piecewise constant $\{-1, 1\}$-valued functions in $[0, T]$, and where f is a vector-valued piecewise analytic function, is compact.

4. Counterexamples. We now show that the assumptions of Theorem 1 cannot be weakened. Clearly, the simplest possible situations in which these assumptions do not hold are:

(a) the control problem

$$\dot{X}(t) = (B + v(t)C)X(t), \tag{3}$$

where B and C are constant matrices such that $BC \neq CB$;

(b) the problem

$$\dot{X}(t) = (u(t)B + v(t)C)X(t), \tag{4}$$

where B and C are as in (a); and

(c) the problem

$$\dot{X}(t) = v(t)F(t)X(t), \tag{5}$$

where F is a matrix-valued function such that $[F(t), F(t')] \neq 0$ for some t, t'.

THEOREM 2. *In each of the cases* (a), (b), (c), *the set* $SB(T)$ *need not be closed.*

Proof. We shall exhibit examples of problems where $SB(T)$ fails to be closed. Our three examples will involve 4×4 matrices, i.e., we shall be working in $GL(4, R)$. The examples for cases (b) and (c) will be derived from the example for case (a).

We let

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Let $X(v, \cdot)$ be the solution of (3) whose value at $t = 0$ is the identity matrix. It is possible to compute $X(v, t)$ explicitly. The result is

$$X(v, t) = \begin{bmatrix} 1 & f(t) & \frac{1}{2}[f(t)]^2 & g(t) \\ 0 & 1 & f(t) & h(t) \\ 0 & 0 & 1 & t \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where

$$f(t) = \int_0^t v(\tau)\, d\tau,$$

$$h(t) = \int_0^t \tau v(\tau)\, d\tau,$$

$$g(t) = \int_0^t v(\tau)h(\tau)\, d\tau.$$

If $SB(T)$ were closed for some $T > 0$, it would follow from Corollary 3 that $SB(T) = S(T)$. In particular, the matrix $X(0, t)$ would belong to $SB(T)$. Thus, there would exist a "bang-bang" control v for which

$$0 = \int_0^T v(\tau)\, d\tau = \int_0^T \tau v(\tau)\, d\tau = \int_0^T v(\tau) \int_0^\tau \lambda v(\lambda)\, d\lambda\, d\tau. \tag{6}$$

We shall show that this is impossible. Indeed, by repeated integrations by parts we obtain:

$$\begin{aligned}\int_0^T v(\tau)\int_0^\tau \lambda v(\lambda)\,d\lambda\,d\tau &= \int_0^T f'(\tau)h(\tau)\,d\tau\\ &= f(T)h(T) - \int_0^T f(\tau)h'(\tau)\,d\tau\\ &= f(T)h(T) - \int_0^T f(\tau)f'(\tau)\tau\,d\tau\\ &= f(T)h(T) - \frac{1}{2}\int_0^T \frac{df^2}{d\tau}(\tau)\tau\,d\tau\\ &= f(T)h(T) - \frac{1}{2}Tf^2(T) + \frac{1}{2}\int_0^T [f(\tau)]^2\,d\tau.\end{aligned}$$

If a control v satisfies (6), it follows that $\int_0^T [f(\tau)]^2\,d\tau = 0$. This implies that $f(\tau) = 0$ almost everywhere. Since v is the derivative of f a.e., we must have that $v = 0$ a.e., so that v cannot be "bang-bang". This completes the proof that $SB(T)$ is not closed for any $T > 0$.

Turning now to case (b), we shall use the same matrices B and C as before. If $X(u, v, \cdot)$ denotes the solution of (4) whose value at $t = 0$ is I, it is clear that $X(1, v, \cdot)$ is the function $X(v, \cdot)$ of the previous paragraph. We claim that $X(1, 0, T)$ cannot be attained by "bang-bang" controls u, v. In view of what we have proved above, it is sufficient to show that, if $X(u, v, T) = X(1, 0, T)$, then $u \equiv 1$. But this can be seen easily as follows: we can compute $X(u, v, T)$ explicitly and obtain $\int_0^T u(\tau)\,d\tau$ as the value of the entry in the third row, fourth column. Since this entry has to be equal to T for $X(1, 0, T)$, and since $u(\tau) \leqq 1$ for all τ, u must be 1 almost everywhere.

Finally, we consider case (c). Here, we define

$$F(t) = \exp(-Bt)\cdot C\cdot \exp(Bt),$$

where B and C are the same matrices that have been used for the other two cases. Let $Y(v, \cdot)$ be the solution of (5) whose value at $t = 0$ is I. It is seen immediately that

$$Y(v, t) = \exp(-Bt)\cdot X(v, t).$$

Since we know that there does not exist a "bang-bang" control v such that $X(v, T) = X(0, T)$, it follows that $Y(0, T)$ is not attainable at time T by "bang-bang" controls. Our proof is thus complete.

Remark 1. It is clear that all our results are equally valid for control problems of the type

$$\dot{x}(t) = [A_0(t) + \sum_{i=1}^{m} u_i(t)A_i(t)]x(t), \tag{7}$$

where $x(t)$ is a column vector in R^n and $A_0, \cdots, A_m$ are matrix-valued functions.[3]

[3] This is the problem studied by Kučera in [5] and [6].

This is obvious for all our positive results, because the solution of (7) with initial condition $x(0) = x_0$ is just $X(u, \cdot)x_0$. As for Theorem 2, we need only observe that the problem considered in case (a) is equivalent to the "vector" problem

$$\dot{x}(t) = (\bar{B} + v(t)\bar{C})x(t),$$

where $x(t)$ is a 4×4 matrix, considered as a vector in 16-dimensional space, and where $\bar{B}$ and $\bar{C}$ are suitable 16×16 matrices (the matrices of the linear transformations $x \to Bx$ and $x \to Cx$, respectively, with respect to an appropriate basis of R^{16}). The sets attainable from the "vector" $x_0 = I$ will coincide with the sets $S(T)$, $SB(T)$, and this way we obtain a counterexample for case (a) of the "vector" problem. Obviously, similar considerations apply to the other cases.

Remark 2. For completeness, we should give an example of a situation in which all the commutativity assumptions of Theorem 1 hold, but $SBP(T)$ fails to be closed because of nonanalyticity of the functions A_i. It is well known (and easy to prove) that the set of numbers $\int_0^T f(t)u(t)\,dt$, where $f(t) = \sin(1/t)$, and where u ranges over all piecewise constant $\{-1, 1\}$-valued functions in $[0, T]$, is not closed for any $T > 0$. Let this be denoted by A_T. The set $B_T = \{e^x : x \in A\}$ is therefore not closed. But B_T is the set of points attainable from $x = 1$ at time T by "bang-bang" controls, for the system

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

(which is of the form that we are considering, with $n = 1$). Moreover, by multiplying the function f by a smooth function that vanishes at the origin to a sufficiently high order, we can modify our counterexample so that f will be as smooth as desired and even C^∞.

Acknowledgment. The author wishes to express his gratitude to R. W. Brockett and V. Jurdjevic for helpful discussions. He also wishes to thank W. W. Xotus, and the referees, for pointing to some inaccuracies in the original manuscript.

REFERENCES

[1] R. W. BROCKETT, *System theory on group manifolds and coset spaces*, submitted to this Journal.

[2] H. HALKIN, *On a generalization of a theorem of Lyapunov*, J. Math. Anal. Appl., 10 (1965), pp. 325–329.

[3] ———, *Some further generalizations of a theorem of Lyapunov*, Arch. Rational Mech. Anal., 17 (1964), pp. 272–277.

[4] G. W. HAYNES AND H. HERMES, *Nonlinear controllability via Lie theory*, this Journal, 7 (1970), pp. 450–460.

[5] J. KUČERA, *Solution in large of control problem* $\dot{x} = (A(1 - u) + Bu)x$, Czech. Math. J., 16 (1966), pp. 600–622.

[6] ———, *Solution in large of control problem* $\dot{x} = (Au + Bv)\dot{x}$, Ibid., 17 (1967), pp. 91–96.

[7] A. LYAPUNOV, *Sur les fonctions-vecteurs complètement additives*, Bull. Acad. Sci. URSS Sér. Math., 4 (1940), pp. 465–478. (In Russian; French summary.)

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

A NOTE ON LEAST SQUARES ESTIMATION BY THE INNOVATIONS METHOD*

THOMAS KAILATH†

Abstract. A rigorous proof by the innovations method is given for certain results on linear least squares estimation, especially of the Stratonovich–Kalman–Bucy formulas. Some discussion is also given of the nonlinear problem.

1. Introduction. In [1]–[2] it was stated that informal arguments were to be used for the derivation of the Stratonovich–Kalman–Bucy least squares estimation formulas by the innovations method. There have been several queries as to a rigorous presentation; we shall provide one such derivation here. It will be seen that the rigorous derivation follows the spirit of the informal one.

The idea of the innovations method is to replace the given observation process by a simpler process that contains the same "information" as the given process. The point is that the estimation problem with the simpler process is usually much easier to solve than the original problem. This method was first used for discrete-time stationary prediction problems by Wold (1938) and Kolmogorov (1939) and then extended to continuous-time stationary processes by Bode and Shannon (1950) and Zadeh and Ragazzini (1950) (these references and several related ones are cited in [1] and [3]). The simpler process has been called the innovations process by Wiener and Masani. For linear least squares problems, we shall use a white-noise process or rather its integrated version, a process with orthogonal increments, as the innovations process; however, other choices are possible (cf. [3], [4]).

In our presentation, we start with the simple but basic problem of finding the least squares estimate of a random variable x, given observations of a related orthogonal-increments process $v(\cdot)$. If we now specialize x to being the state of a linear dynamical system and if we take $v(\cdot)$ as the innovations process of noisy observations of the state of this system, then we can readily obtain the Stratonovich–Kalman–Bucy equations [5, p. 675], [6]. Finally in § 4 we note that by using certain recent results the innovations method can also be extended to certain nonlinear problems.

2. Linear estimation of a random variable from an orthogonal-increments process. Let x be a random variable with

$$Ex = 0, \quad E|x|^2 < \infty, \tag{1}$$

and let $\{v(\tau), 0 \leqq \tau \leqq t\}$ be a related process of orthogonal increments with

$$Ev(t) = 0, \quad Ev(t)v'(s) = \int_0^{t \wedge s} I \, d\tau, \tag{2}$$

where $t \wedge s$ denotes the minimum of t and s and the prime denotes transpose.

* Received by the editors July 19, 1971, and in revised form October 8, 1971.

† Department of Electrical Engineering, Stanford University, Stanford, California 94305. This work was supported by the Applied Mathematics Division of the Air Force Office of Scientific Research under Contract AF44-620-69-C-0101 and by the JSEP at Stanford University, Contract N-00014-67-A-0112-0044.

We could have assumed that the covariance of $v(\cdot)$ was the integral of a positive-definite function, but the normalization to the identity function is convenient and entails no loss of generality. We wish to find the least squares estimate, $\hat{x}(t)$, in the form

$$\hat{x}(t) = \int_0^t g_t(s)\, dv(s), \tag{3}$$

where

$$\int_0^t \operatorname{tr}\,[g_t(s)g_t'(s)]\, ds < \infty. \tag{4}$$

There is no loss of generality here, since the so-called Wiener stochastic integral in (3) is the most general form of a finite-variance linear functional of the process $\{v(\tau), 0 \leqq \tau \leqq t\}$ (see, for example, Doob [7]).

To determine the weighting function $g_t(s)$, we shall use the well-known criterion that

$$\tilde{x}(t) \triangleq x - \hat{x}(t) \perp v(\tau), \qquad 0 \leqq \tau \leqq t, \tag{5}$$

where $\perp$ denotes orthogonality (zero-correlation). This yields

$$Exv'(\tau) = E\left[\int_0^t g_t(s)\, dv(s) \int_0^\tau 1' \cdot dv'(u)\right] = \int_0^\tau g_t(s)\, ds \tag{6}$$

by using the properties of Wiener stochastic integrals [7, Chap. IX]. Therefore,

$$g_t(\tau) = \frac{\partial}{\partial \tau} Exv'(\tau), \qquad \tau \leqq t, \tag{7}$$

and

$$\hat{x}(t) = \int_0^t \left[\frac{\partial}{\partial \tau} Exv'(\tau)\right] dv(\tau). \tag{8}$$

To go further, we need more assumptions on x. A common set of assumptions is treated in the next section.

3. The Stratonovich–Kalman–Bucy problem. We now assume that

$$x = x(t), \tag{9}$$

where $x(\cdot)$ is the solution of a linear stochastic differential equation

$$dx(\tau) = F(\tau)x(\tau)\, d\tau + G(\tau)\, dU(\tau), \quad x(0) = x_0, \tag{10}$$

and $U(\cdot)$ is a process of orthogonal increments such that

$$EU(\tau) = 0, \quad EU(t)U'(s) = \int_0^{t \wedge s} Q(\tau)\, d\tau, \qquad Q \geqq 0, \tag{11}$$

and

$$Ex_0 = 0, \quad Ex_0x_0' = \Pi_0, \quad EU(t)x_0' = 0, \qquad t \geqq 0. \tag{12}$$

We also assume that we have observations of the form

$$dY(\tau) = z(\tau)\,d\tau + dW(\tau), \quad z(\tau) = H(\tau)x(\tau), \tag{13}$$

where $W(\cdot)$ is an orthogonal-increments process with

$$EW(t) = 0, \quad EW(t)W'(s) = \int_0^{t\wedge s} I\,d\tau, \tag{14}$$

$$EW(t)x_0' \equiv 0, \quad EW(t)U'(s) = \int_0^{t\wedge s} C(\tau)\,d\tau. \tag{15}$$

We shall also assume that

$$F(\cdot), G(\cdot), H(\cdot) \quad \text{are continuous functions.} \tag{16}$$

As will be seen from the proofs below, the assumptions (16) are stronger than really necessary, but they are the conventional ones.

We may remark that in derivations using white-noise and delta functions, the presence of the correlation term $C(\cdot)$ usually leads to various unsatisfactory discussions about "partial" delta functions and to the use of relations like (see [1], [6])

$$\int_{-\infty}^{t} \delta(t-\tau)\,d\tau = \tfrac{1}{2}.$$

In the author's opinion, the need to clarify this problem is perhaps a more important motivation for going to a stochastic integral treatment than just the desire to obtain a formal mathematical proof.

Next, we introduce the fundamental matrix of $F(\cdot)$, which is defined as the unique solution of the equation

$$d\Phi(t,s)/dt = F(t)\Phi(t,s), \quad \Phi(0,0) = I.$$

In terms of $\Phi(\cdot,\cdot)$, we can write the solution $x(\cdot)$ of (10) as

$$x(t) = \Phi(t,0)x_0 + \int_0^t \Phi(t,s)G(s)\,dU(s). \tag{17}$$

We also note the properties

$$E[W(t) - W(s)]x'(\tau) = 0, \qquad \tau \leqq s,t, \tag{18}$$

and

$$E\int_0^T \operatorname{tr}\,[H(s)x(s)x'(s)H'(s)]\,ds < \infty. \tag{19}$$

(a) *The innovations process.* For the solution, the first step is to obtain the innovations process. From (13), (18), (19) it can be shown that the process

$$\upsilon(t) = Y(t) - \int_0^t H(s)\hat{x}(s)\,ds = \int_0^t H(s)\tilde{x}(s)\,ds + W(t) \tag{20}$$

is a process of orthogonal increments with

$$E\upsilon(t) = 0, \quad E\upsilon(t)\upsilon'(s) = \int_0^{t\wedge s} I\, d\tau \tag{21}$$

and that

$$\sigma\{\upsilon(\tau), \tau \leqq t\} = \sigma\{Y(\tau), \tau \leqq t\}, \qquad 0 \leqq t \leqq T, \tag{22}$$

where $\sigma\{\cdot\}$ denotes the sigma fields generated by the specified random variables. Therefore, $\upsilon(\cdot)$ is the innovations process of the observations $Y(\cdot)$. Proofs of the fundamental equations (20)–(22) are outlined in the Appendix.

Next, in order to calculate $\hat{x}(t)$, we shall need (cf. (8)) to evaluate

$$g_t(s) = \frac{\partial}{\partial s} E[x(t)\upsilon'(s)].$$

We shall find, as will be shown presently, that the assumptions (10)–(15) on $x(\cdot)$ easily yield

$$g_t(s) = \Phi(t, s)K(s), \qquad 0 \leqq s \leqq t, \tag{23}$$

where

$$K(t) = P(t)H'(t) + G(t)C'(t) \tag{24}$$

and

$$P(t) = E\tilde{x}(t)\tilde{x}'(t), \quad \tilde{x}(t) = x(t) - \hat{x}(t). \tag{25}$$

Therefore, by (8),

$$\hat{x}(t) = \int_0^t \Phi(t, s)K(s)\, d\upsilon(s), \tag{26}$$

or in differential form,

$$d\hat{x}(t) = F(t)\hat{x}(t)\, dt + K(t)\, d\upsilon(t), \quad \hat{x}(0) = 0, \tag{27}$$

which is the celebrated equation of Stratonovich, Kalman and Bucy.

Proof of (23)–(24). We have, by (20),

$$\begin{aligned} g_t(s) &= \frac{\partial}{\partial s} E[x(t)\upsilon'(s)] \qquad (s \leqq t) \\ &= \frac{\partial}{\partial s} E\left[\int_0^s x(t)\tilde{x}'(u)H'(u)\, du\right] + \frac{\partial}{\partial s} E[x(t)W'(s)]. \end{aligned} \tag{28}$$

By the use of Schwarz' inequality and the Fubini theorem, we can show that the first term on the right-hand side of (28) is

$$\begin{aligned} \frac{\partial}{\partial s}\int_0^s E[x(t)\tilde{x}'(u)H'(u)]\, du &= E[x(t)\tilde{x}'(s)]H'(s) \\ &= \Phi(t, s)E[x(s)\tilde{x}'(s)]H'(s) \\ &\quad + E\left[\int_s^t \Phi(t, \tau)G(\tau)\, dU(\tau)\cdot \tilde{x}'(s)\right]H'(s) \\ &= \Phi(t, s)P(s)H'(s) + 0, \end{aligned}$$

while the second term is equal to

$$\frac{\partial}{\partial s}E\left[\Phi(t,0)x_0W'(s) + \int_0^t \Phi(t,\tau)G(\tau)\,dU(\tau)\int_0^s dW(\sigma)\right]$$

$$= \frac{\partial}{\partial s}\int_0^s \Phi(t,\tau)G(\tau)C(\tau)\,d\tau = \Phi(t,s)G(s)C(s), \qquad s \leqq t.$$

The sum of these two terms yields the desired relations (23)–(24).

(b) *Derivation of Riccati equation for* $P(\cdot)$. The mean square error $P(t)$ obeys a differential equation, which is also easily obtained via the innovations. We first observe that for the process $x(\cdot)$ described by the stochastic differential equation (10)–(12), a direct calculation using (17) yields

$$\Pi(t) \triangleq Ex(t)x'(t) = \Phi(t,0)\Pi_0\Phi'(t,0) + \int_0^t \Phi(t,\tau)G(\tau)Q(\tau)G'(\tau)\Phi'(t,\tau)\,d\tau.$$

Equivalently, we have the differential equation

$$\dot{\Pi}(t) = F(t)\Pi(t) + \Pi(t)F'(t) + G(t)Q(t)G'(t), \quad \Pi(0) = \Pi_0. \tag{29}$$

Since $v(\cdot)$ has orthogonal increments, a similar calculation applied to $\hat{x}(\cdot)$, as given by the stochastic differential equation (27), yields

$$\Sigma(t) \triangleq E\hat{x}(t)\hat{x}'(t), \quad \dot{\Sigma}(t) = F(t)\Sigma(t) + \Sigma(t)F'(t) + K(t)IK'(t), \quad \Sigma(0) = 0. \tag{30}$$

But since $\hat{x}(\cdot)$ and $\tilde{x}(\cdot)$ are orthogonal, we have

$$\Pi(t) = Ex(t)x'(t) = E\hat{x}(t)\hat{x}'(t) + E\tilde{x}(t)\tilde{x}'(t) = \Sigma(t) + P(t), \tag{31}$$

and therefore,

$$\begin{aligned} \dot{P}(t) &= \dot{\Pi}(t) - \dot{\Sigma}(t) \\ &= F(t)P(t) + P(t)F'(t) - K(t)K'(t) + G(t)Q(t)G'(t), \\ P(0) &= \Pi_0. \end{aligned} \tag{32}$$

This is the well-known matrix Riccati differential equation of the Stratonovich–Kalman–Bucy theory; the above derivation was first given in [3, footnote 7].

As far as the author is aware, the only previous direct rigorous proof of the Stratonovich–Kalman–Bucy result is the one due to Wonham [8], which proceeds via a careful limiting argument from the discrete-time case. Some rigorous proofs have been proposed via specializations of certain nonlinear filtering formulas, but until recently all such proofs have needed fairly strong conditions on the process $x(\cdot)$. Recently, however, Fujisaki, Kallianpur and Kunita [9] have given a derivation (partly based on innovations) under fairly general conditions; their results can be specialized to the linear case, but such a proof will be less direct than the one we have given here. Even more recently, Balakrishnan [23] has given a proof based on martingale arguments, but it does not seem to be as direct as ours.

4. Some extensions; nonlinear estimation. Similar arguments can be used to justify the results obtained by the innovations method for linear smoothing [2], and linear filtering and smoothing based only on covariance functions [10], [11];

various extensions, for example, to infinite-dimensional and distributed parameter systems, can also be made.

Somewhat surprisingly the innovations method can also be used [12] for nonlinear least squares estimation given observations of the form

$$dY(t) = z(t)\,dt + dW(t),$$

where $W(\cdot)$ is a Wiener process but $z(\cdot)$ is not necessarily Gaussian. If

$$E\int |z(t)|\,dt < \infty$$

and the increments of $W(\cdot)$ are independent of the sigma fields generated by past $z(\cdot)$ and $W(\cdot)$, then it can be shown [13], [14] that $v(\cdot)$ defined by

$$dv = dY - \hat{z}(t)\,dt, \quad \hat{z}(t) = E[z(t)|\sigma\{Y(\tau), \tau \leqq t\}]$$

is a Wiener process with the same covariance function as $W(\cdot)$. The difficult problem is to show that there is no loss of information in working with $v(\cdot)$, i.e., $\sigma\{v(\tau), \tau \leqq t\} = \sigma\{Y(\tau), \tau \leqq t\}$. In [12], the rather strong hypothesis

$$|z(t)| \leqq M$$

was necessary. Recently, however, Yerzhov [15] has shown that a sufficient condition for the equality of the sigma fields associated with $v(\cdot)$ and $Y(\cdot)$ is that the measures induced in the space of continuous functions by the processes $Y(\cdot)$ and $W(\cdot)$ be mutually absolutely continuous. This will be the case, for instance, in the widely studied problem in which it is assumed that $z(\cdot)$ and $W(\cdot)$ are completely independent and $z(\cdot)$ is square-integrable almost surely. (Several other assumptions will also yield mutual absolute continuity.) Given the innovations process, a derivation very similar to that of the linear problem can be used (as informally described in [12], see also [9]) to obtain the differential equations of Stratonovich, Kushner and others for the conditional density function. Here we shall only quote the analogue of the linear estimation result in § 2. [*Note added in proof*. Shepp and P. Varaiya have pointed out to me that Yerzhov's proof appears to be inadequate.]

The least squares estimate of a finite-variance random variable x given observations $\{v(\tau), 0 \leqq \tau \leqq t\}$ of a Wiener process $v(\cdot)$ can be written

$$\begin{aligned} \hat{x}(t) &= E[x(t)|\,\mathfrak{Y}_t], \\ &= \int_0^t \left\{ \frac{\partial}{\partial \tau} E[xv'(\tau)|\,\mathfrak{Y}_\tau] \right\} dv(\tau), \qquad \mathfrak{Y}_t = \sigma\{v(\tau), \tau \leqq t\}. \end{aligned} \tag{33}$$

The analogy to (8) is striking. If we now make further assumptions on x, for example, that x is the value at time t of a diffusion process or a Markov jump process, etc., then from the representation (33), we can obtain equations for the evolution of $\hat{x}(t)$ or more generally of the conditional characteristic function or the conditional probability density function.

Appendix: the innovations results of (20)–(22). There are many ways to prove that the process $v(\cdot)$ defined by (20) of the main text is a process of orthogonal increments. For example, one could specialize the more general result of [13], [14], quoted in §4, which is obtained by using certain martingale theorems. Or one could

use the simpler arguments of (61)–(62) of [16]. It is somewhat more difficult to show that the process $Y(\cdot)$ can be uniquely recovered from the process $v(\cdot)$. Hitsuda [17] has used martingale theory to give a rigorous derivation of (20)–(22). However, here we shall outline a proof based on a factorization theorem of Gohberg and Krein [18]; this proof is the rigorous counterpart of the informal arguments used in [1] and [3]. The result of Gohberg and Krein is for operators on certain general Hilbert spaces, but here we shall need only the result for the space L_2.

Let K be an integral operator on $L_2[0,1]$, with a real symmetric square-integrable kernel (for notational simplicity, we assume everything scalar in this Appendix)

$$K(t,s) = K(s,t), \quad \int_0^1 \int_0^1 K^2(t,s)\,dt\,ds < \infty. \tag{A.1}$$

Suppose also that the operator $I + K$ is positive definite, or equivalently that the eigenvalues of K are all greater than -1. Then Gohberg and Krein [18] show that the equation

$$h(t,s) + \int_0^t h(t,\tau)K(\tau,s)\,d\tau = K(t,s), \qquad 0 \leqq s \leqq t \leqq 1, \tag{A.2}$$

$$h(t,s) = 0, \qquad s > t, \tag{A.3}$$

has a unique square-integrable solution and that

$$(I + K)^{-1} = (I - h^*)(I - h), \tag{A.4}$$

where h^* denotes the adjoint of the Volterra operator h that has kernel $h(t,s)$.

Furthermore, if $k(t,s)$ is the unique square-integrable solution of

$$h(t,s) + \int_0^t k(t,\tau)h(\tau,s)\,d\tau = k(t,s), \qquad 0 \leqq s \leqq t \leqq 1, \tag{A.5}$$

$$k(t,s) = 0, \qquad s > t, \tag{A.6}$$

then we can also write

$$I + K = (I + k)(I + k^*). \tag{A.7}$$

We use these results as follows. Suppose that

$$dY = z(t)\,dt + dW,$$

where

$$EW(t)W(s) = \int_0^{t\wedge s} d\tau, \quad E\int_0^1 z^2(t)\,dt < \infty,$$

and the increments of $W(\cdot)$ are uncorrelated with past values of $z(\cdot)$. It is easy to verify that the assumptions (10)–(16) of the main text yield a process $Y(\cdot)$ that satisfies these conditions. The first thing to do is to show that a process $Y(\cdot)$ defined as above has covariance function

$$EY(t)Y(s) = t \wedge s + \int_0^t \int_0^s K(u,v)\,du\,dv,$$

where $K(u, v)$ meets the conditions of the Gohberg–Krein theorem. This fact can be established by a direct calculation—see, for example, the discussion of (48)–(49) in [19]. (A more elegant method is to assume temporarily that $z(\cdot)$ and $W(\cdot)$ are Gaussian processes and then to note that, by any one of a variety of criteria, the process $Y(\cdot)$ is mutually absolutely continuous with respect to the Wiener process $W(\cdot)$. Therefore, by a theorem of Shepp [20] (see also [21], [22]), the covariance of Y must have the stated properties.)

The next step is to observe that the linear least squares estimate of $z(\cdot)$ can be written as

$$\hat{z}(t) = \int_0^t h(t, s)\, dY(s) = \int_0^t h(t, s) z(s)\, ds + \int_0^t h(t, s)\, dW(s),$$

where $h(t, s)$ is defined by (A.2)–(A.3). This follows easily by using the projection definition of $\hat{z}(\cdot)$. Now the fact that

$$v(t) = Y(t) - \int_0^t d\tau \int_0^\tau h(\tau, u)\, dY(u)$$

has covariance function $t \wedge s$ follows by a direct but tedious calculation. The basic idea is symbolically that

$$\begin{aligned} \dot{v}_t &= \dot{Y}_t - \int_0^t h(t, u)\, dY(u) \\ &= [(I - h)Y]_t \end{aligned}$$

so that

$$\begin{aligned} E\dot{v}_t\dot{v}_s &= (I - h)E\dot{Y}_t\dot{Y}_s(I - h^*) \\ &= (I - h)(I + k)(I + k^*)(I - h^*) = I. \end{aligned}$$

Finally, since $(I - h)$ is invertible, we expect that $Y(\cdot)$ can be recovered from $v(\cdot)$ by

$$\dot{Y}_t = (I - h)^{-1}\dot{v}_t = (I + k)\dot{v}_t$$

or equivalently,

$$Y(t) = v(t) + \int_0^t d\tau \int_0^\tau k(\tau, u)\, dv(u).$$

This last equation can be directly verified without much difficulty. The right-hand side is equal to

$$Y(t) - \int_0^t d\tau \int_0^\tau h(\tau, u)\, dY(u) + \int_0^t d\tau \int_0^\tau k(\tau, u)\, dY(u)$$

$$- \int_0^t d\tau \int_0^\tau k(\tau, u) \int_0^u h(u, \sigma)\, dY(\sigma)\, du.$$

By the use of a Fubini theorem for Wiener integrals [7, pp. 430–431], the last term can be rewritten as

$$\int_0^t d\tau \int_0^\tau dY(\sigma) \int_0^\tau k(\tau, u)h(u, \sigma)\, du.$$

But now the sum of the last three terms is

$$-\int_0^t d\tau \left[\int_0^\tau \left[h(\tau, \sigma) - k(\tau, \sigma) + \int_0^\tau k(\tau, u)h(u, \sigma)\, du \right] dY(\sigma) \right],$$

which is zero because the term in square brackets is zero almost everywhere by the definition of h and k.

Acknowledgment. It is a pleasure to acknowledge discussions on § 2 with G. Kallianpur.

REFERENCES

[1] T. KAILATH, *An innovations approach to least-squares estimation, Part I: Linear filtering in additive white noise*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 646–655.

[2] T. KAILATH AND P. FROST, *An innovations approach to least-squares estimation, Part II: Linear smoothing in additive white noise*, Ibid., AC-13 (1968), pp. 655–660.

[3] T. KAILATH, *The innovations approach to detection and estimation theory*, Proc. IEEE, 58 (1970), pp. 680–695.

[4] D. L. DUTTWEILER, *Reproducing kernel Hilbert space techniques for detection and estimation problems*, Doctoral thesis, Department of Electrical Engineering, Stanford University, Stanford, Calif., 1970.

[5] R. L. STRATONOVICH, *Detection and estimation of signals in non-Gaussian noise*, Proc. IEEE, 58 (1970), pp. 670–679.

[6] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME J. Basic Engrg., Ser. A, 83 (1961), pp. 95–107.

[7] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[8] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. T. Bharucha-Reid, ed., Academic Press, New York, 1969, pp. 131–212.

[9] M. FUJISAKI, G. KALLIANPUR AND H. KUNITA, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., to appear.

[10] T. KAILATH AND R. GEESEY, *An innovations approach to least-squares estimation, Part IV: Recursive estimation given the covariance function*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 720–727.

[11] R. GEESEY AND T. KAILATH, *An innovations approach to least-squares estimation, Part V: Recursive estimation in colored noise*, submitted for publication.

[12] P. FROST AND T. KAILATH, *An innovations approach to least-squares estimation, Part III: Nonlinear least-squares estimation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 217–226.

[13] T. KAILATH, *Some extensions of the innovations theorem*, Bell System Tech. J., 50 (1971), pp. 1487–1494.

[14] R. LIPTSER AND A. SHIRYAEV, *On absolutely continuous measures, corresponding to processes of diffusion type, relative to Wiener measure*, Izv. Akad. Nauk., to appear.

[15] M. YERZHOV, *Existence and uniqueness of solutions of stochastic functional equations*, Teor. Veroyatnost. i Primenen., to appear.

[16] T. KAILATH, *A general likelihood-ratio formula for random signals*, IEEE Trans. Information Theory, IT-15 (1969), pp. 350–361.

[17] M. HITSUDA, *Representation of Gaussian processes equivalent to Wiener processes*, Osaka J. Math., 5 (1968), pp. 299–312.

[18] I. Ts. Gohberg and M. G. Krein, *The factorization of operators in Hilbert space*, Amer. Math. Soc. Transl. (2), 51 (1966), pp. 155–188.
[19] T. Kailath, *Likelihood ratios for Gaussian processes*, IEEE Trans. Information Theory, IT-16 (1970), pp. 276–288.
[20] L. Shepp, *Radon–Nikodym derivatives of Gaussian measures*, Ann. Math. Statist., 37 (1966), pp. 321–354.
[21] T. Kailath, *On measures equivalent to Wiener measure*, Ibid., 38 (1967), pp. 261–263.
[22] G. Kallianpur and H. Oodaira, *The equivalence and singularity of Gaussian measures*, Proc. Symp. on Time Series Analysis, M. Rosenblatt, ed., John Wiley, New York, 1963, pp. 000–000.
[23] A. V. Balakrishnan, *Stochastic differential systems, I*, Lecture notes, Dept. of System Science, UCLA, 1971.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

OPTIMAL CONTROLS FOR PROBLEMS WITH A RESTRICTED STATE SPACE*

A. B. SCHWARZKOPF†

Abstract. In this paper we derive conditions necessary for a curve to minimize a functional of the Bolza form, subject to differential equation side conditions and the restriction that the trajectory of an admissible curve is constrained to lie inside a closed set S with a smooth boundary. These results are obtained as limits of conditions derived by E. J. McShane for curves whose trajectories are contained in a neighborhood of S. This limit procedure verifies the continuity of optimal solutions as a "soft" boundary of S "hardens" until it allows no penetration at all.

1. Introduction. Consider the system consisting of equations

$$x^i(t) = x^i(a) + \int_a^t f^i(\gamma, x(\gamma), u(\gamma))\, d\gamma, \qquad i = 1, 2, \cdots, n,$$

and the cost functional

$$J(u) = \int_a^b f^0(\gamma, x(\gamma), u(\gamma))\, d\gamma + e(a, x(a), b, x(b))$$

governed by the control $u(t)$ which takes values in the set $U \subseteq R^m$. E. J. McShane [1] has established conditions which this system must satisfy if J is to be minimized over the collection of all curves obtained by using relaxed control functions defined on a bounded set U (or on an unbounded set under additional assumptions about the rate of growth of the control). McShane's results assume that the trajectory $x(t)$ lies in the interior of some phase space in R^n, and in this paper we extend these results to give conditions which must hold when U is bounded but the trajectories are allowed to pass along the boundary of the phase space S. Conditions similar to those we present in § 3 appear in various places in the literature of optimal control and the calculus of variations. For example, [2, Theorem 22], [3], [6], [7, § 8], [8, § 8], [9], [10] and [11, Chap. 8] all contain similar results under varying hypotheses. The work of Chang [6] and Warga [3] comes closest to that presented here in that they both consider the bounded state variable problem as a limit of the unbounded case. Chang's paper gives an intuitive derivation of a part of our Theorem 2, using a penalty function similar to the one we introduce in § 4.1b. Warga on the other hand proves a slight generalization of our Theorems 2 and 3 using a slightly different definition of relaxed control, and intermediate constraints which force the approximating optimal trajectories to remain within the set S at specified points of time.

The other references mentioned above obtain necessary conditions for an optimal curve directly and do not demonstrate the fact that the optimal curve in this case is the limit of optimal curves where the set S has a soft boundary. The work of Neustadt [7], [8] and of Dubovitsky and Milyutin [9] is particularly

* Received by the editors May 5, 1970, and in final revised form May 3, 1971.

† Department of Mathematics, U.S. Naval Postgraduate School, Monterey, California. Now at Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73069. Some of this work was done as part of a doctoral dissertation under E. J. McShane at the University of Virginia, where it was supported in part by the Army Research Office under Grant ARO-D-31-124-G662.

interesting since these references obtain their results as a special case of abstract necessary conditions which apply to a variety of problems. Chapter 8 of [7] also contains a history of the optimization problem with bounded state variables.

The integrals in the following work are ordinary Lebesgue integrals with the exception of our discussion of relaxed controls where we use Lebesgue–Stieltjes integration. A more elementary treatment of this topic is available in [1, § 1]. From time to time we misuse the integral notation to the extent that if a function ϕ is defined on a subset M of the interval $[a, b]$, and M has measure $|b - a|$, then we will understand that

$$\int_a^b \phi(\gamma)\, d\gamma = \int_M \phi(\gamma)\, d\gamma$$

whenever the right-hand integral exists, even though the integrand on the left may not be defined for some values in $[a, b]$.

2. Relaxed controls and generalized curves. We begin by defining relaxed control functions and generalized curves. Further discussions of these concepts are available in [1], [3], and [4].

2.1. Relaxed controls. Suppose U is a closed and bounded subset of n-dimensional real space R^n and $[a, b]$ is a finite real interval.

DEFINITION 2.1a. A *relaxed control* function is a function which assigns to each t in $[a, b]$ a mean value process $\mathfrak{M}[\,\cdot\,; t]$ over U such that if $\phi(t, u)$ is a bounded continuous function defined on $[a, b] \times U$, then the function $\mathfrak{M}[\phi(t, u); t]$ is Lebesgue measurable. In other words if $a \leqq t \leqq b$, then $\mathfrak{M}[\phi(t, u); t] \equiv \int_U \phi(t, u)\, d\mu$ where μ is a probability measure on U.

Throughout this paper we assume that all relaxed control functions are defined on the whole real line by setting $\mathfrak{M}[\,\cdot\,; t] = 0$ where it is not otherwise defined, although we must keep in mind that $\mathfrak{M}$ is not a mean value operator except for t in the interval on which it is defined by Definition 2.1a.

If $\mathfrak{M}$ is a relaxed control function, we define the *support of* $\mathfrak{M}$ *at* t_0 to be the set of points u_0 in U such that if ϕ is a bounded, continuous, nonnegative function defined on U and ϕ is positive at u_0 then $\mathfrak{M}[\phi(u), t_0] > 0$. We define the *general support of* $\mathfrak{M}$ similarly to be the set of all points (t_0, u_0) in $R \times U$ such that if $\phi(t, u)$ is a bounded, continuous, nonnegative function defined on $R \times U$ which is positive at (t_0, u_0), then $\int_{-\infty}^{\infty} \mathfrak{M}[\phi(t, u); t]\, dt > 0$.

2.2. Generalized curves. We define a *generalized curve* C to be a triple $(\mathfrak{M}, x(\,\cdot\,), [a, b])$ such that $\mathfrak{M}$ is a relaxed control function which is zero outside of the interval $[a, b]$ and $x(\,\cdot\,)$ is the n-dimensional trajectory defined by the equations

$$x^i(t) = x^i(a) + \int_a^t \mathfrak{M}[f^i(\gamma, x(\gamma), u); \gamma]\, d\gamma, \qquad \begin{matrix} i = 1, 2, \cdots, n, \\ \alpha \leqq t \leqq \beta. \end{matrix} \tag{2.2a}$$

The functions $f^i(t, x, u)$, $t \in R$, $x \in R^n$, $u \in U$ are predetermined by the optimization problem, and assumed to be bounded and continuous real-valued functions. We remark that if the relaxed control function assigns unit measure to a point $u_0(t)$

for almost all t in $[a, b]$, then (2.2a) becomes

$$x^i(t) = x^i(a) + \int_a^t f^i(\gamma, x(\gamma), u_0(\gamma))\, d\gamma, \qquad i = 1, 2, \cdots, n,$$

and the generalized curve $C = (\mathfrak{M}, x(\cdot), [a, b])$ is in fact an ordinary curve. For a further discussion of the relationship between generalized and ordinary curves see [1, § 10].

2.3. Limits of generalized curves. Let us suppose that the functions $f^i(t, x, u)$, $i = 1, 2, \cdots, n$, are bounded, continuous real-valued functions defined on $T \times B \times U$, where

B is a closed subset of R^n called the state space,

U is a closed and bounded subset of R^m called the control set,

$T = [\alpha, \beta]$ is a closed finite interval in R,

E is a closed set of points in R^{2n+2} of the form (t_0, x_0, t_1, x_1), where x_0 and x_1 are in B, and t_0 and t_1 are in T. E is called the set of endpoints.

The collection G of *admissible* curves is defined to be the collection of all generalized curves $C = (\mathfrak{M}, x(t), [a, b])$ such that:

(2.3a) The trajectory $x(t) = (x^1(t), x^2(t), \cdots, x^n(t))$ is governed by the set of integral equations

$$x^i(t) = x^i(a) + \int_a^t \mathfrak{M}[f^i(\gamma, x(\gamma), u); \gamma]\, d\gamma, \qquad \alpha \leqq t \leqq \beta, \quad i = 1, 2, \cdots, n.$$

(2.3b) The interval $[a, b]$ is contained in T; the point $(a, x(a), b, x(b)) \equiv \eta(C)$ is in E; the set of points $\{x(t) | a \leqq t \leqq b\}$ is in B; and the support of $\mathfrak{M}[\cdot\,; t]$ is contained in U for every t in $[a, b]$.

The collection of all admissible generalized curves above satisfies a weak compactness condition under the following definition of covergence. We say the sequence $\{C_q = (\mathfrak{M}_q, x_q, [a_q, b_q])\}_{q=1}^{\infty}$ of admissible generalized curves *converges* to a generalized curve $C_0 = (\mathfrak{M}_0, x_0, [a_0, b_0])$ if:

(2.3c) (i) $\lim_{q \to \infty} a_q = a_0$ and $\lim_{q \to \infty} b_q = b_0$,

(ii) $\lim_{q \to \infty} x_q(t) = x_0(t)$ for all t in $T = [\alpha, \beta]$,

(iii) for every bounded and continuous function ϕ with compact support defined on $T \times R^n \times U$,

$$\lim_{q \to \infty} \int_\alpha^t \mathfrak{M}_q[\phi(\gamma, x_q(\gamma), u); \gamma]\, d\gamma = \int_\alpha^t \mathfrak{M}_0[\phi(\gamma, x_0(\gamma), u); \gamma]\, d\gamma$$

for all t in T.

Recall that we have defined relaxed controls so that if $C = (\mathfrak{M}, x, [a, b])$ is a generalized curve, then $\mathfrak{M}$ is zero outside the interval $[a, b]$ and

$$\int_\alpha^t \mathfrak{M}[\phi(\gamma, x(\gamma), u); \gamma]\, d\gamma = \int_a^t \mathfrak{M}[\phi(\gamma, x(\gamma), u); \gamma]\, d\gamma$$

whenever either integral is defined.

THEOREM 2.3d. *Let G be the collection of admissible generalized curves defined above and assume the notation and hypotheses set forth there. Suppose further that G is not empty, and that there exists a sequence $\{C_q\}_{q=1}^{\infty}$ of generalized curves from G whose trajectories all lie in a closed (in R^n) and bounded subset A of B. Then*:

(i) *There is a generalized curve $C_0 = (\mathfrak{M}_0, x_0, [a_0, b_0])$ in G and a subsequence Γ of the positive integers such that C_0 is the limit of the curves C_q as $q \to \infty$, q in Γ. Moreover the trajectory x_0 is the uniform limit of the curves x_q on T.*

(ii) *There exists a subset M of T such that M has Lebesgue measure $|\beta - \alpha|$ and for t_0 in M and ϕ any bounded, continuous function with compact support defined on $T \times R^n \times U$,*

$$\mathfrak{M}_0[\phi(t_0, x_0(t_0), u); t_0] = \frac{d}{dt}\int_{\alpha}^{t} \mathfrak{M}_0[\phi(\gamma, x_0(\gamma), u); \gamma]\, d\gamma\Big|_{t=t_0}.$$

(iii) *If u_0 is a point in the support $\mathfrak{M}_0$ at t_0 (in M), then there is a sequence of pairs (t_{q_j}, u_{q_j}) with u_{q_j} in the support of $\mathfrak{M}_{q_j}$ at t_{q_j} such that*

$$\lim_{q_j \to \infty} (t_{q_j}, u_{q_j}) = (t_0, u_0).$$

Moreover the numbers may be chosen to miss any subset A of measure zero in T such that $(T - A) \supseteq M$.

Proof. Results (i) and (ii) follow from a careful reading of Theorem 2.7 in [1], so we prove only assertion (iii). Let A be a subset of T with measure zero, let $\operatorname{supp} \mathfrak{M}[\,\cdot\,; t]$ denote the support of the relaxed control at the point t, and let $S(\mathfrak{M})$ denote the general support of $\mathfrak{M}$. We shall first show that every point in $S(\mathfrak{M}_0)$ can be expressed as the limit of points (t_{q_i}, u_{q_i}), where $u_{q_i} \in \operatorname{supp} \mathfrak{M}_{q_i}[\,\cdot\,; t_{q_i}]$ and $t_{q_i} \notin A$. Suppose that this is false. Then there exists a point (t_0, u_0) in $S(\mathfrak{M}_0)$, a number q_0, and a ball N of radius ε about (t_0, u_0) such that for $q > q_0$ there is no point (t, u) in N with $t \notin A$ and $u \in \operatorname{supp} \mathfrak{M}_q[\,\cdot\,; t]$. Let $\phi(t, u)$ be a nonnegative, continuous, real-valued function on $T \times U$ which is zero outside of N and positive at (t_0, u_0). Since ϕ is zero on $\operatorname{supp} \mathfrak{M}_q[\,\cdot\,; t]$ for almost all t in T and $q > q_0$ we must have $\mathfrak{M}_q[\phi(t, u); t] = 0$ a.e. if $q > q_0$, and hence

$$\lim_{q_j \to \infty} \int_{\alpha}^{\beta} \mathfrak{M}_{q_j}[\phi(t, u); t]\, dt = 0.$$

This is impossible since (i) and (ii) imply that

$$\int_{\alpha}^{\beta} \mathfrak{M}_0[\phi(t, u); t]\, dt = \lim_{q_j \to \infty} \int_{\alpha}^{\beta} \mathfrak{M}_{q_j}[\phi(t, u); t]\, dt = 0,$$

and the left-hand integral must be positive since $(t_0, u_0) \in S(\mathfrak{M}_0)$. We conclude that every point in $S(\mathfrak{M}_0)$ can be written as claimed.

We conclude our proof by showing that

$$S(\mathfrak{M}_0) \supseteq \bigcup_{t \in M} (\{t\} \times [\operatorname{supp} \mathfrak{M}[\,\cdot\,; t]]),$$

where M is defined in Theorem 2.3d (ii). Suppose $t_0 \in M$ and $u_0 \in \operatorname{supp} \mathfrak{M}_0[\,\cdot\,; t_0]$. Then for any nonnegative continuous function $\psi(t, u)$ defined on $T \times U$ which is

positive at (t_0, u_0) we have by (ii),

$$\frac{d}{dt}\int_\alpha^t \mathfrak{M}_0[\psi(\gamma, u); \gamma]\, d\gamma|_{t=t_0} = \mathfrak{M}_0[\psi(t_0, u); t_0] > 0.$$

Since the function

$$\Psi(t) = \int_\alpha^t \mathfrak{M}_0[\psi(\gamma, u); \gamma]\, d\gamma$$

is monotone increasing with a positive derivative at $t = t_0$, we must have $\Psi(\beta) > 0$ and (t_0, u_0) is in $S(\mathfrak{M}_0)$. The theorem follows by choosing A to contain $T - M$.

3. Optimal generalized curves. Let us call an admissible generalized curve *S-admissible* provided its trajectory $x(t)$ remains inside the set $S = \{x : p(x(t)) \leqq 0\}$, where $p(x)$ is a continuous function defined on R^n in such a way that $S \subseteq B$. In the sequel we refer to the set $S' = R^n - S$ as the *forbidden set*. In this section we shall state conditions necessary for an S-admissible curve to minimize the functional

$$J(C) = \int_a^b \mathfrak{M}[f^0(\gamma, x(\gamma), u); \gamma]\, d\gamma + e(\eta(C)),$$

where $f^0(t, x, u)$ is a continuous function defined on $T \times B \times U$ and $e(t_0, x_0, t_1, x_1)$ is a continuous function defined on E. Any such optimal curve will be called *S-optimal.*

If the set S above is bounded, then E. J. McShane has shown that an optimal curve, i.e., one that minimizes J over the class of S-admissible curves, does exist [1, Theorem 2.7]. His theorem is essentially Theorem 2.3d of the previous section applied to the functional J. In addition McShane presents a set of necessary conditions for an optimal curve which does not touch the boundary of S. The following results generalize his results to the problem posed above.

3.1. Notation and assumptions. The assumptions of § 2.3 were sufficient to prove the existence of an S-optimal curve, but we must make some additional assumptions in order to derive conditions which further specify such curves.

(3.1a) Suppose that each function $f^i(t, x, u)$, $i = 1, \cdots, n$, defined in § 2.3 is bounded and continuous, along with its first and second order partial derivatives.

(3.1b) Suppose the function $p(x)$ defined above is bounded and continuous along with its first and second order partial derivatives. Moreover suppose that there is a number $\zeta > 0$ such that the set $\{x : x \in R^n, p(x) \leqq \zeta\}$ is a bounded subset of B.

(3.1c) Suppose that the function $e(t_0, x_0, t_1, x_1)$ defined on E is continuous and has continuous partial derivatives.

Let η denote a point in E, and suppose that there exists a continuously differentiable function L_η which maps R^{2n+2} into itself with $L_\eta(0) = \eta$.

(3.1d) Further suppose that there exists a convex cone K containing 0 such that $L_\eta(K) \subseteq E$. Now define the set E'_η to be the set of all vectors in R^{2n+2} of the form $L'_\eta(0; h), h$ in K. Here $L'_\eta(0; h)$ denotes the differential of L_η evaluated at 0 and in the direction h. We shall denote the set E'_η by E', in spite of the fact that it depends on the particular point η chosen in E. The set E' is a convex cone which we call the *cone of inward directions to E at η*.

Before stating Theorem 3.2a, which is the main result of this paper, we should define two functions which considerably simplify notation. Suppose λ is a point in R^{n+1}, and let $t \in T$, $x \in B$, and $u \in U$. We define

$$F(t, x, u, \lambda) = \sum_{i=0}^{n} \lambda^i f^i(t, x, u)$$

and

$$b(t, x, u) = \sum_{i=1}^{n} p_{x^i}(x) \cdot f^i(t, x, u).$$

We are now prepared to state Theorem 3.2a.

3.2. Necessary conditions for an S-optimal generalized curve.

THEOREM 3.2a. *Assume the hypotheses and notation of* §§ 2.3 *and* 3.1. *Suppose that the collection of S-admissible curves defined in* § 3 *is not empty. Then there exists an S-admissible generalized curve* $C_0 = (\mathfrak{M}_0, x_0, [a_0, b_0])$ *which gives the functional*

$$J(C) = \int_a^b \mathfrak{M}_0[f^0(s, x(s), u); s]\, ds + e(a, x(a), b, x(b))$$

a value equal to its infimum over all S-admissible curves. Moreover, the curve C_0 *may be chosen to satisfy the conditions below.*

(i) *There exist multipliers* $\lambda^0, \lambda^1, \cdots, \lambda^n, \ell$ *defined on* $[a_0, b_0]$ *such that* λ^0 *is a nonnegative constant, and the function* $\ell(t)$ *is nonnegative, monotonely decreasing, and constant on any interval on which* $p(x_0(t))$ *is negative. Moreover* $\ell(t)$ *satisfies the inequalities*

$$1 \geqq \ell(t) \geqq \ell(b_0) = 0.$$

(ii) *The functions* $\lambda^1, \lambda^2, \cdots, \lambda^n$ *satisfy the integral equations*

$$\lambda^i(t) = \lambda^i(a_0) - \int_{a_0}^{t} \mathfrak{M}_0[F_{x^i}(s, x_0(s), u, \lambda(s)); s]\, ds$$

$$- \int_{a_0}^{t} \mathfrak{M}_0[b_{x^i}(s, x_0(s), u) \cdot \ell(s); s]\, ds$$

for $i = 1, 2, \cdots, n$ *and* $a_0 \leqq t \leqq b_0$.

(iii) *There exists a finite constant c such that if*

$$\mu(t) = c + \int_{a_0}^{t} \mathfrak{M}_0[F_t(s, x_0(s), u, \lambda(s)); s]\, ds$$

$$+ \int_{a_0}^{t} \mathfrak{M}_0[b_t(s, x_0(s), u) \cdot \ell(s); s]\, ds,$$

then

$$\inf_{u\in U}\{F(t,x_0(t),u,\lambda(t))+b(t,x_0(t),u)\cdot\ell(t)\}=\mu(t)$$

almost everywhere in $[a_0,b_0]$, *and in particular equality holds wherever* ℓ *is continuous.*

(iv) *There exists a subset* A' *of* $[a_0,b_0]$ *whose Lebesgue measure is zero, such that for* t *in* $([a_0,b_0]-A')$ *and* u_0 *in the support of* $\mathfrak{M}_0$ *at* t *we have*

$$\mu(t)=F(t,x_0(t),u_0,\lambda(t))+b(t,x_0(t),u_0)\cdot\ell(t).$$

(v) *Let* $(\gamma_0,\xi_0,\gamma_1,\xi_1)$ *be any vector in the set of interior directions* E' *at the endpoint* $\eta(C_0)$, *where* γ_0 *and* γ_1 *are in* R^1 *and* ξ_0 *and* ξ_1 *are in* R^n. *Then*

$$\left\{\sum_{j=1}^{n}\left[\lambda^0\frac{\partial e}{\partial x_0^j}+\lambda^j(a_0)+\ell(a_0)\cdot p_{x^j}(x_0(a_0))\right]\xi_0^j+\sum_{j=1}^{n}\left[\lambda^0\frac{\partial e}{\partial x_1^j}-\lambda^j(b_0)\right]\xi_1^j\right.$$
$$\left.+\left[\lambda^0\frac{\partial e}{\partial t_0}-\mu(a_0)\right]\gamma_0+\left[\lambda^0\frac{\partial e}{\partial t_0}+\mu(b_0)\right]\gamma_1\right\}\geqq 0,$$

where the partial derivatives are evaluated at $\eta(C_0)$.

3.3. Further results. In working with results like Theorem 3.2a, it is usually most helpful to know that the multipliers involved do not vanish. We have not shown that this is true in general, but the following result, suggested by Warga's result [3, (3.1.2.8)], does give a reasonable set of conditions under which these functions do not all vanish at any point of T. Suppose $C=(\mathfrak{M},x(t),[a,b])$ is an S optimal curve satisfying Theorem 3.2a.

THEOREM 3.3a. *If there exists a positive number* v *such that*

$$\inf_{u\in U}b(t,x(t),u)\leqq -v<0$$

for all t *such that* $p(x(t))=0$, *then either the functions* $\lambda^0,\lambda^1,\cdots,\lambda^n,\ell$ *are all zero everywhere in* $(a,b]$, *or they do not all vanish anywhere in* $[a,b]$. *Moreover, if* $p(x(a))<0$, *then the first alternative is impossible.*

Theorems 3.2a and 3.3a give a fairly comprehensive set of necessary conditions for one S-optimal curve, but we have not shown that these conditions hold for every S-optimal curve. The following result largely overcomes this problem.

THEOREM 3.3b. *Corresponding to any S-optimal curve* C', *there is an S-optimal curve* C *which has the same trajectory as* C', *and satisfies the conclusions of Theorem* 3.2a *and Theorem* 3.3a.

4. Proof of Theorem 3.2a.

4.1. The q-approximation problem.

Our basic approach to the proof of Theorem 3.2a is to replace the stated problem with an approximation to it, which satisfies the hypotheses of McShane's work [1]. Our approximation allows curves to enter the forbidden set S' but penalizes such an encroachment. Let H denote the collection of all admissible curves defined in § 2 and suppose that $C=(\mathfrak{M},x(\cdot),[a,b])$ is a curve in H. We wish to adjoin two additional coordinates, y^1 and y^2, to the trajectory x of this curve, subject to the conditions (4.1a) and (4.1b).

The function $y^1(t)$ is defined by the equation

$$y^1(t) = p(x(a)) + \int_\alpha^t \mathfrak{M}[b(\gamma, x(\gamma), u); \gamma]\, d\gamma. \tag{4.1a}$$

Choose some function ϕ which is continuously differentiable and monotonely increasing on the entire real line, and satisfies the relations

$$\phi(z)\begin{cases} = 0 & \text{for } z \leqq -1, \\ > 0 & \text{for } -1 < z < 0, \\ = 1 & \text{for } 0 \leqq z. \end{cases}$$

Define $\Phi(z)$ by

$$\Phi(z) = \int_0^z \phi(\gamma)\, d\gamma$$

so that

$$\begin{aligned} -1/q \leqq \Phi(q \cdot z)/q \leqq 0 \quad & \text{for } z \leqq 0, \\ \Phi(q \cdot z)/q = z \quad & \text{for } z \geqq 0. \end{aligned}$$

Notice that the derivative of $\Phi(q \cdot z)/q$ with respect to q is positive for $-1/q \leqq z \leqq 0$ and is zero otherwise, so $\Phi(q \cdot z)/q$ is monotonely increasing in q. Now for a given $q > 0$, define the function $y^2(q, \cdot)$ by the equation

$$y^2(q, t) = \int_\alpha \mathfrak{M}[\Phi(q \cdot y^1(\gamma))/q; \gamma]\, d\gamma, \tag{4.1b}$$

and observe that $y^2(q, t)$ is also monotonely increasing in q for any given generalized curve C and fixed t.

We shall use the term *q-admissible* to refer to any curve $C^* = (\mathfrak{M}, (x(\cdot), y^1(\cdot), y^2(q \cdot)), [a, b])$ obtained by adjoining the functions $y^1(\cdot)$ and $y^2(q, \cdot)$ to a curve C in H, if C^* satisfies the additional requirement.

$$y^2(q, b) \leqq 0. \tag{4.1c}$$

Denote the collection of all q-admissible curves C^* by the symbol H_q^*. We pose a new optimization problem for curves in H_q^* which approximates the original problem of this theorem.

The q-approximation problem. From the collection of all curves C^* in H_q^*, find those, if any, for which the functional

$$J(C^*) = \int_a^b \mathfrak{M}[f^0(\gamma, x(\gamma), u); \gamma]\, d\gamma + e(\eta(C^*)) \tag{4.1d}$$

attains the minimum value possible in H_q^*.

We shall call any solution to the q-approximation problem above a *q-optimal* curve.

We defined y^1 so that $y^1(t) = p(x(t))$. Suppose $p(x(t_0)) > 0$. Then if (4.1c) is satisfied, there must be points at which $p(x(t))$ is negative. By (3.1a) and (3.1b), $p(x(t))$ is Lipschitz with some constant D so

$$p(x(t_0)) - D|t - t_0| \leqq p(x(t)).$$

The left-hand side of this inequality is positive whenever t is in either $[t_0 - (p(x(t_0))/D), t_0]$ or $[t_0, t_0 + (p(x(t_0))/D)]$, and since $p(x(t))$ is zero at some point of $[a, b]$ at least one of these intervals is contained in $[a, b]$. Integrating over the appropriate interval and observing that $[p(x(t))]^+ \geqq 0$ gives the inequality

$$\int_a^b [p(x(t)]^+ \, dt \geqq \frac{p(x(t_0))^2}{2D}.$$

On the other hand,

$$\begin{aligned}\int_a^b [p(x(t))]^+ \, dt &= \int_a^b [\Phi(q \cdot p(x(t)))/q]^+ \, dt \\ &\leqq \int_a^b [\Phi(q \cdot p(x(t)))/q]^- \, dt \\ &\leqq [b - a]/q \leqq [\beta - \alpha]/q.\end{aligned}$$

Combining the two relations above gives

$$\frac{p(x(t_0))^2}{2D} \leqq \frac{[\beta - \alpha]}{q}$$

or

$$p(x(t_0)) \leqq \left|\frac{2D[\beta - \alpha]}{q}\right|^{1/2} \equiv Kq^{-1/2} \tag{4.1e}$$

Thus all q-admissible curves have $y^2(q, b) \leqq K \cdot q^{-1/2}$, where K is independent of q. Moreover since $y^2(q, \beta)$ is increasing in q, the sets H_q^* are nested and shrinking.

4.2. Solution of the q-approximate problem. In this section we wish to find a curve C_q^*, which is optimal for the corresponding q-approximate problem if q is sufficiently large. To do this we shall appeal to the results obtained by McShane [1]. We observe first that the collection H_q^* is not empty, since any curve C^* corresponding to a curve C, which is S-admissible for the original problem, is in H_q^* for every positive q. For sufficiently large q, formula (4.1e) shows that every curve C^* in H_q^* satisfies the inequality $y^1(t) = p(x(t)) < \zeta$ for every t in $[a, b]$. Since the set $\{x : p(x) < \zeta\}$ is bounded by assumption, the trajectory $x(t)$ is contained in a single bounded set independent of the particular choice of C^* from H_q^*. It follows that the functions $y^1(t)$, and $y^2(q, t)$ are also uniformly bounded for C^* in H_q^*, and the trajectory $(x(t), y^1(t), y^2(q, t))$ remains in a bounded subset of R^{n+2} for all such curves C^*. We now apply one of McShane's results [1, Theorem 2.7], or equivalently Theorem 2.3d stated above, to show that for all sufficiently large q, there exists a corresponding q-optimal curve C_q^*.

Suppose next that $C_q^* = (\mathfrak{M}_q, (x_q(t), y_q^1(t), y_q^2(q,t)), [a_q, b_q])$ is any such q-optimal curve. The q-approximate problem satisfies the hypotheses of Theorem 4.7 in [1] so there must be multipliers $\lambda_q^0, \lambda_q^1, \cdots, \lambda_q^n, \ell_q^1, \ell_q^2$ defined on $[a_q, b_q]$ which satisfy the conclusions of that theorem. The results of the remainder of this section are all obtained by applying these conclusions to the particular curve C_q^* and the corresponding multipliers above. *It will simplify our notation considerably if we drop the subscript q throughout this section, and we shall do so.* Our present notation will be resumed in the next section. According to McShane's theorem the multiplier λ^0 is either 0 or 1 and the function $\ell^2(t)$ must be a constant. The other multipliers satisfy the differential equations

$$\lambda^i(t) = \lambda^i(a) - \int_a^t \mathfrak{M}[F_{x^i}(\gamma, x(\gamma), u, \lambda(\gamma)) + b_{x^i}(\gamma, x(\gamma), u) \cdot \ell^1(\gamma); \gamma]\, d\gamma,$$

(4.2a)

$$\ell^1(t) = \ell^1(a) - \int_a^t \phi(q \cdot y^1(\gamma)) \cdot \ell^2\, d\gamma, \qquad i = 1, 2, \cdots, n.$$

We note further that the functions $\lambda^0, \lambda^1, \cdots, \lambda^n, \ell^1, \ell^2$ are not all zero at any point of $[a, b]$. Furthermore if we define the function $\mu(t)$ by

$$\mu(t) = \inf_{u \in U} \{F(t, x(t), u, \lambda(t)) + b(t, x(t), u) \cdot \ell^1(t) + \ell^2 \cdot \Phi(q \cdot y^1(t))/q\},$$

then there exists a finite constant c such that

$$\mu(t) = c + \int_a^t \mathfrak{M}[F_t(\gamma, x(\gamma), u, \lambda(\gamma)) + b_t(\gamma, x(\gamma), u) \cdot \ell^1(\gamma); \gamma]\, d\gamma \tag{4.2b}$$

for all t in $[a, b]$.

For almost all t in $[a, b]$ and for all u_0 in the support of $\mathfrak{M}$ at t,

$$\mu(t) = \{F(t, x(t), u_0, \lambda(t)) + b(t, x(t), u_0) \cdot \ell^1(t) + \ell^2 \cdot \Phi(qy^1(t))/q\}. \tag{4.2c}$$

If E^* denotes the set of allowable endpoints for the q-approximate problem we are considering, we can define the set $(E^*)'$ of inward directions at a point $(t_0, x_0, y_0^1, y_0^2, t_1, x_1, y_1^1, y_1^2)$ in E^* to be the set of all vectors $(\gamma_0, \xi_0, w_0^1, w_0^2, \gamma_1, \xi_1, w_1^1, w_1^2)$ such that $(\gamma_0, \xi_0, \gamma_1, \xi_1)$ is in the set E' at (t_0, x_0, t_1, x_1) and the following conditions hold:

$$w_0^1 = \sum_{k=1}^n \xi_0^k \cdot p_{x^k}(x_0), \quad w_0^2 = 0, \quad w_1^1 \in R,$$

$$w_1^2 \begin{cases} \in R & \text{if } y^2(q, b) < 0, \\ \leqq 0 & \text{if } y^2(q, b) = 0. \end{cases}$$

Under these circumstances the last result of McShane's theorem is

$$\begin{aligned} &\sum_{j=1}^n \left[\lambda^0 \cdot \frac{\partial e}{\partial x_0^j} + \lambda^j(a)\right] \cdot \xi_0^j + \ell^1(a) \cdot w_0^1 \\ &\quad + \sum_{j=1}^n \left[\lambda^0 \cdot \frac{\partial e}{\partial x_1^j} - \lambda^j(b)\right] \cdot \xi_1^j - \ell^1(b) \cdot w_1^1 - \ell^2(b) \cdot w_1^2 \\ &\quad + \left[\lambda^0 \cdot \frac{\partial e}{\partial t_0} - \mu(a)\right] \cdot \gamma_0 + \left[\lambda^0 \cdot \frac{\partial e}{\partial t_1} + \mu(b)\right] \cdot \gamma_1 \geqq 0, \end{aligned} \tag{4.2d}$$

and all partial derivatives are evaluated at $(a, x(a), b, x(b))$.

Let us consider these conditions further. The vectors $(0, 0, \cdots, 0, w_1^1, 0)$ and $-(0, 0, \cdots, 0, w_1^1, 0)$ are both in E^* so (4.2d) implies that $\ell^1(b) = 0$. Furthermore if $y^2(q, b) < 0$, then a similar argument shows that ℓ^2 must be zero at b, and since ℓ^2 is constant, it must be zero everywhere. A quick look at (4.2a) shows that ℓ^1 is constant if ℓ^2 is zero, and we have shown that $\ell^1(b)$ must be zero. In other words, if $y^2(q, b) < 0$, then ℓ^1 and ℓ^2 are identically zero. On the other hand, if $y^2(q, b) = 0$ and $\ell^2 \neq 0$, then $w_1^2 \leqq 0$ and hence $\ell^2 > 0$. It follows that $\ell^1(t)$ is a monotone decreasing function which is zero at $t = b$. More precisely, $\ell^1(t)$ is constant on precisely those intervals for which $y^1(t) \leqq -1/q$. Since we are assuming that $y^2(q, b) = \int_a^b [\Phi(q \cdot y^1(\gamma))/q]\, d\gamma = 0$ and since $\Phi(q \cdot y^1(t))$ is strictly negative when $y^1(t) < 0$, there must be some interval on which $y^1(t) > -1/q$. Hence $\ell^1(t)$ is not identically zero. According to [1] the functions $\lambda^0(t), \lambda^1(t), \cdots, \lambda^n(t), \ell^1(t), \ell^2(t)$ do not all vanish at any point in $[a, b]$ and in particular they do not all vanish at $t = a$. We have already shown that if $\ell^2 \neq 0$, then $\ell^1(a) > 0$; and, hence, at least one of the functions $\lambda^0(t), \lambda^1(t), \cdots, \lambda^n(t), \ell^1(t)$ must be nonzero for $t = a$. *We shall assume therefore that*

$$(\ell^1(a))^2 + \sum_{j=0}^{n} (\lambda^j(a))^2 = 1. \tag{4.2e}$$

For future reference we shall summarize our conclusions about the function $\ell^1(t)$.

> The function $\ell^1(t)$ is defined and monotonely decreasing on the interval $[a, b]$ with
>
> $$1 \geqq \ell^1(t) \geqq \ell^1(b) = 0. \tag{4.2f}$$
>
> Moreover $\ell^1(t)$ is constant on precisely those intervals on which $p(x(t)) \leqq -1/q$.

We conclude this section with a lemma.

LEMMA 4.2g. *For t and s in the interval $[a, b]$,*

$$\sum_{j=1}^{n} (\lambda^j(t))^2 \leqq e^{K_1|t-s|}\left[\sum_{j=1}^{n} (\lambda^j(s))^2\right] + K_2\left|\int_s^t (\ell^1(\gamma))^2\, d\gamma\right| + K_2 \cdot (\lambda^0)^2|t - s|,$$

where the constants K_1 and K_2 are independent of q.

Proof. Let the function $\beta(t)$ be defined by the left-hand side of the above inequality, and let K_0 be a bound for the functions $f^k_{x^i}$ and b_{x^i} for all $k, i = 1, 2, \cdots, n$. Hypotheses (3.1a) and (3.1b) assure that such a bound exists. Thus for almost all t in $[a, b]$ we have

$$\begin{aligned}
\frac{d}{dt}\beta(t)\bigg|_t &= \sum_{j=1}^{n} -2\lambda^j(t)\{\mathfrak{M}[F_{x^j}(t, x(t), u, \lambda(t)) + \ell^1(t) \cdot b_{x^j}(t, x(t), u); t]\} \\
&= \sum_{j=1}^{n} -2\lambda^j(t)\left\{\sum_{k=0}^{n} \lambda^k(t) \cdot \mathfrak{M}[f^k_{x^j}(t, x(t), u); t] + \ell^1(t) \cdot \mathfrak{M}[b_{x^j}(t, x(t), u); t]\right\} \\
&\leqq \sum_{j=1}^{n} 2|\lambda^j(t)|\left\{\sum_{k=0}^{n} |\lambda^k(t)|K_0 + \ell^1(t)K_0\right\} \qquad \text{(cont.)}
\end{aligned}$$

$$\leqq 2K_0 \cdot \sum_{j=1}^{n} \left\{ \sum_{k=1}^{n} \left[\sum_{h=1}^{n} ((\lambda^h(t)))^2 \right] + |\lambda^j(t)|\lambda^0 + |\lambda^j(t)|\ell^1(t) \right\}$$

$$\leqq 2K_0 \cdot \sum_{j=1}^{n} \{n \cdot (\beta(t)) + (\lambda^j(t))^2 + (\lambda^0)^2 + (\lambda^j(t))^2 + (\ell^1(t))^2\}$$

$$\leqq 2K_0 \cdot \sum_{j=1}^{n} \{n\beta(t) + \beta(t) + (\lambda^0)^2 + \beta(t) + (\ell^1(t))^2\}$$

$$= 2K_0 n(n+2) \cdot \beta(t) + 2K_0 n((\lambda^0)^2 + (\ell^1(t))^2).$$

Now we apply Grönwall's lemma and see

$$\beta(t) \leqq e^{K_1|t-s|}\beta(s) + e^{K_1|t-s|}\left| \int_s^t e^{-K_1|\gamma-s|} 2nK_0((\lambda^0)^2 + \ell^1(\gamma)^2)\, d\gamma \right|$$

$$\leqq e^{K_1|t-s|}\left\{ \beta(s) + 2nK_0 \left| \int_s^t (\lambda^0)^2 + (\ell^1(\gamma))^2\, d\gamma \right| \right\}$$

which is the required formula when

$$K_1 = 2K_0 n(n+2) \quad \text{and} \quad K_2 = 2nK_0\, e^{K_1|\beta-\alpha|}.$$

4.3. Passing to a limit. In § 4.2 we showed that there must exist a q-optimal curve C_q^* for any large q, and we have given conditions which each of these curves must satisfy. In this section we shall show that some subsequence of the sequence $(C_q^*,\ q = 1, 2, 3, \cdots)$ converges to a limit C_0^*, and that this limit corresponds to a curve which is S-optimal.

We begin by choosing one particular optimal curve C_q^* for each $q = 1, 2, 3, \cdots$. This can be done for large integers, and since we are interested in finding a limit as q increases, we shall ignore the fact that C_q^* may not exist for small q. Each curve $C_q^* = (\mathfrak{M}_q, x_q(t), y_q^1(t), y_q^2(q,t), [a_q, b_q])$ corresponds to a curve $C_q = (\mathfrak{M}_q, x_q(t), [a_q, b_q])$ which is admissible but possibly not S admissible (§ 3). The arguments at the beginning of § 4.2 show that for large q, the trajectories $x_q(t)$ are all contained in the bounded set $\{x : p(x) < \zeta\}$, so we apply Theorem 2.3d which guarantees a subsequence Γ of the positive integers and an admissible curve $C_0 = (\mathfrak{M}_0, x_0(t), [a_0, b_0])$ such that

$$\lim_{q \to \infty} C_q = C_0, \qquad q \in \Gamma, \tag{4.3a}$$

in the sense of (2.3a). A glance at inequality (4.1e) plus a little calculation reveals that

$$\begin{aligned} y_0^1(t) &= p(x_0(\alpha)) + \int_\alpha^t \mathfrak{M}_0[b(\gamma, x_0(\gamma), u);\ \gamma]\, d\gamma \\ &= p(x_0(t)) \leqq 0, \end{aligned} \tag{4.3b}$$

and C_0 has a trajectory which remains inside of S. If we adjoin coordinates $y^1(t)$ and $y^2(q,t)$ to any S-admissible curve D, we obtain a curve D_q^* which is q-admissible,

and hence $J(C_q^*) \leqq J(D_q^*) = J(D)$. On the other hand, Theorem 2.3d says that

$$\begin{aligned}\lim_{q\to\infty} J(C_q^*) &= \lim_{q\to\infty}\left[\int_\alpha^\beta \mathfrak{M}_q[f^0(\gamma, x_q(\gamma), u); \gamma]\, d\gamma + e(\eta(C_q^*))\right] \\ &= \int_\alpha^\beta \mathfrak{M}_0[f^0(\gamma, x_0(\gamma), u); \gamma]\, d\gamma + e(\eta(C_q^*)) \\ &= J(C_0^*) = J(C_0),\end{aligned}$$

and hence C_0 must minimize the functional J over the class of all S-admissible curves.

4.4. The minimum principle: Conditions (i)–(iv). In the previous section we showed that there must exist an S-optimal curve C_0, and that C_0 can be obtained as the limit of curves C_q, each of which corresponds to a q-optimal curve C_q^*. Furthermore, in § 4.2 we showed that for each C_q^* there exist multipliers $\lambda_q^0(t)$, $\lambda_q^1(t), \cdots, \lambda_q^n(t), \ell_q^1(t), \ell_q^2(t)$ which satisfy conditions (4.2a) through Lemma 4.2g. In this section we obtain results (i) through (iv) of Theorem 2.3d as limiting cases of these formulas.

In the following calculations recall that we could have defined the trajectories $x_q(t), y_q^1(t), y_q^2(q, t)$, and $x_0(t)$ as well as the multipliers defined in § 4.2 on the entire interval $T = [\alpha, \beta]$ by setting them equal to their values at a_q and b_q on the intervals $[\alpha, a_q)$ and $(b_q, \beta]$ respectively. This is consistent with the convention of Definition 2.1a already adopted which defines $\mathfrak{M}_q[\,\cdot\,; t]$ to be constantly zero on these intervals. Thus (4.2e) says that the vectors $(\lambda_q^0(\alpha), \lambda_q^1(\alpha), \cdots, \lambda_q^n(\alpha), \ell_q^1(\alpha))$ are all contained in a closed unit ball, and by compactness we may choose Γ so that the sequence of vectors $(\lambda_q^0(\alpha), \lambda_q^1(\alpha), \cdots, \lambda_q^n(\alpha), \ell_q^1(\alpha))$ approaches the limit $(\lambda_0^0(\alpha), \lambda_0^1(\alpha), \cdots, \lambda_0^n(\alpha), \ell_0^1(\alpha))$ as q increases in Γ. Since the functions $\lambda_q^0(t)$ are all nonnegative constants, the function $\lambda_0^0(t)$ defined by the limit

$$\lambda_0^0(t) = \lim_{q\to\infty} \lambda_q^0(\alpha), \qquad q \in \Gamma, \quad \alpha \leqq t \leqq \beta,$$

is also a nonnegative constant. Furthermore, since the functions $\ell_q^1(t)$ are all uniformly bounded and monotone on Γ, we use Helley's selection theorem to see that there is a monotone decreasing point function $\ell_0^1(t)$ defined for $\alpha \leqq t \leqq \beta$ such that

$$\ell_0^1(t) = \lim_{q\to\infty} \ell_q^1(t), \qquad q \in \Gamma. \tag{4.4a}$$

Since the functions $\ell_q^1(t)$ are constant on those intervals on which $p(x_q(t)) \leqq -1/q$, it follows that $\ell_0^1(t)$ must be constant on those intervals for which $p(x_0(t))$ is negative. This proves (i) of Theorem 3.2a. Furthermore, since $\ell_0^1(t)$ is monotone and bounded on $[\alpha, \beta]$ it must be continuous except at countably many points of $[\alpha, \beta]$. We also observe that if $p(x_0(\alpha)) < 0$, then $(\lambda_0^0, \lambda_0^1(a_0), \cdots, \lambda_0^n(a_0), \ell_0(a_0))$ is not a zero vector. This is true since all the components are continuous at a_0, and have the same value at α as at a_0.

Next we observe that Lemma 4.2g and condition (4.2e) say that the functions $\lambda_q^i(t)$, $i = 1, 2, \cdots, n$, q in Γ, are uniformly bounded on T, and a careful inspection of the right-hand side of (4.2a) shows that the integrands of these functions are

uniformly bounded. We now use Ascoli's theorem to define functions $\lambda_0^i(t)$ by

$$\lambda_0^i(t) = \lim_{q\to\infty} \lambda_q^i(t), \qquad q\in\Gamma, \quad t\in\Gamma.$$

Next we note that

$$\begin{aligned}
&\left|\int_\alpha^\beta \mathfrak{M}_q[\ell_q^1(\gamma)\cdot b_{x^i}(\gamma, x_q(\gamma), u);\gamma]\,d\gamma - \int_\alpha^\beta \mathfrak{M}_0[\ell_0^1(\gamma)\cdot b_{x^i}(\gamma, x_0(\gamma), u);\gamma]\,d\gamma\right| \\
&\qquad \leqq \left|\int_\alpha^\beta [\ell_q^1(\gamma) - \ell_0^1(\gamma)]\cdot \mathfrak{M}_q[b_{x^i}(\gamma, x_q(\gamma), u);\gamma]\,d\gamma\right| \\
&\qquad\quad + \left|\int_\alpha^\beta \ell_0^1(\gamma)\cdot\{\mathfrak{M}_q[b_{x^i}(\gamma, x_q(\gamma), u);\gamma]\right. \\
&\qquad\qquad \left. - \mathfrak{M}_0(b_{x^i}[(\gamma, x_0(\gamma), u);\gamma]\}\,d\gamma\right|.
\end{aligned}$$

Using dominated convergence on the first term, and (2.3c(iii)) on the second term on the right above, we see that both terms approach zero as q increases in Γ. This and similar arguments show that

$$\begin{aligned}
\lambda_0^i(t) &= \lambda_0^i(\alpha) - \lim_{q\to\infty}\int_\alpha^t \mathfrak{M}_q[F_{x^i}(\gamma, x_q(\gamma), u, \lambda_q(\gamma)) + \ell_q^1(\gamma)\cdot b_{x^i}(\gamma, x_q(\gamma), u);\gamma]\,d\gamma \\
&= \lambda_0^i(\alpha) - \lim_{q\to\infty}\int_\alpha^t \mathfrak{M}_q\left[\sum_{k=0}^n \lambda^k(\gamma)\cdot f^k_{x^i}(\gamma, x_q(\gamma), u);\gamma\right]d\gamma \\
&\quad - \lim_{q\to\infty}\int_\alpha^t \mathfrak{M}_q[\ell_q^1(\gamma)\cdot b_{x^i}(\gamma, x_q(\gamma), u);\gamma]\,d\gamma \\
&= \lambda_0^i(\alpha) - \int_\alpha^t \mathfrak{M}_0[F_{x^i}(\gamma, x_0(\gamma), u, \lambda_0(\gamma)) + \ell_0^1(\gamma)\cdot b_{x^i}(\gamma, x_0(\gamma), u);\gamma]\,d\gamma.
\end{aligned} \tag{4.4b}$$

This is the formula (ii) of Theorem 3.2a.

We show next that $\ell_q^2\cdot\Phi(qy_q^1(t))/q$ approaches zero almost everywhere in $[a_0, b_0]$, after which condition (iii) of Theorem 3.2a follows from (4.2b) as a limit calculation similar to the one above.

LEMMA 4.4c. *If $\ell_0^1(t)$ is continuous at $t = t_0$, then*

$$\lim_{q\to\infty} \ell_q^2\Phi(qy_q^1(t_0))/q = 0, \qquad q\in\Gamma.$$

Proof. We begin by showing that for large q,

$$\int_{t_0}^{t_0+h} \ell_q^2\Phi(q\cdot y_q^1(\gamma))/q\,d\gamma = y_q^2(q, t_0+h) - y_q^2(q, t_0)$$

is small. We have assumed that $y_q^2(q,\beta)\geqq 0$ in (4.1c), and in the argument following (4.2d) we showed that if $y_q^2(q,\beta) < 0$, then $\ell_q^2 = 0$ and our result follows. Suppose

$y_q^2(q, \beta) = 0$. Then

$$\int_{a_q}^{b_q} [\ell_q^2 \Phi(q y_q^1(\gamma))/q]^- \, d\gamma = \int_{a_q}^{b_q} [\ell_q^2 \Phi(q y_q^1(\gamma))/q]^+ \, d\gamma$$

$$= \int_{a_q}^{b_q} [\ell_q^2 y_q^1(\gamma)]^+ \, d\gamma \tag{4.4d}$$

$$\leqq \ell_q^2 \cdot \sup_{a_q \leqq t \leqq b_q} y_q^1(t) \cdot \|A_q\|,$$

where $\|A_q\|$ denotes the Lebesgue measure of the set $A_q = \{t : y_q^1(t) \geqq 0\}$. Furthermore,

$$1 \geqq \ell_q^1(a_q) - \ell_q^1(b_q) = \int_{b_q}^{a_q} \ell_q^2 \phi(q y_q^1(\gamma)) \, d\gamma \geqq \ell_q^2 \cdot \|A_q\|.$$

Solving the last inequality for ℓ_q^2 and substituting into (4.4d) gives

$$\int_{t_0}^{t_0+h} |\ell_q^2 \Phi(q y_q^1(\gamma))/q| \, d\gamma \leqq \int_{a_q}^{b_q} |\ell_q^2 \Phi(q y_q^1(\gamma))/q| \, d\gamma$$

$$= 2 \int_{a_q}^{b_q} [\ell_q^2 \Phi(q y_q^1(\gamma))/q]^+ \, d\gamma$$

$$\leqq \frac{2}{\|A_q\|} \cdot \sup_{a_q \leqq t \leqq b_q} y_q^1(t) \cdot \|A_q\|$$

$$\leqq 2Kq^{-1/2}$$

by (4.1e). On the other hand, we can apply Taylor's theorem with integral remainder to write

$$\int_{t_0}^{t_0+h} \ell_q^2 \Phi(q y_q^1(\gamma))/q \, d\gamma = h \cdot \ell_q^2 \Phi(q y_q^1(t_0))/q + \int_{t_0}^{t_0+h} (t_0 + h - \gamma) \ell_q^2 \phi(q y_q^1(\gamma)) \mathfrak{M}_q[b(\gamma, x_q(\gamma), u); \gamma] \, d\gamma.$$

Rearranging and dividing by h gives

$$\left| \frac{\ell_q^2 \Phi(q y_q^1(t_0))}{q} \right| \leqq \left| \int_{t_0}^{t_0+h} \ell_q^2 \Phi(q y_q^1(\gamma))/q \, d\gamma \right| \cdot \frac{1}{|h|} + \frac{1}{|h|} \left| \int_{t_0}^{t_0+h} (t_0 + h - \gamma) \ell_q^2 \phi(q y_q^1(\gamma)) \mathfrak{M}_q[b(\gamma, x_q(\gamma), u); \gamma] \, d\gamma \right|$$

$$\leqq \frac{K}{|h| \cdot q^{1/2}} + \frac{1}{|h|} \cdot |t_0 + h - t_0| \sup_{t_0 \leqq t \leqq t_0+h} |\mathfrak{M}_q[b(\gamma, x_q(\gamma), u); t]| \cdot \left| \int_{t_0}^{t_0+h} \ell_q^2 \phi(q y_q^1(\gamma)) \, d\gamma \right|$$

$$\leqq \frac{K}{|h| q^{1/2}} + D|\ell_q^1(t_0 + h) - \ell_q^1(t_0)|,$$

where D is a bound for $b(t, x, u)$ as assured by hypotheses (3.1a) and (3.1b) and the definition of b. Now let $\varepsilon > 0$ be given. Since ℓ_0^1 is continuous at $t = t_0$ we may choose some fixed h such that

$$|\ell_0^1(t_0 + h) - \ell_0^1(t_0)| < \varepsilon/4D.$$

Furthermore, since $\ell_q^1(t)$ converges pointwise to $\ell_0^1(t)$, we may choose Q such that for $q > Q$ the following three inequalities hold:

$$|\ell_q^1(t_0) - \ell_0^1(t_0)| < \varepsilon/4D,$$
$$|\ell_q^1(t_0 + h) - \ell_0^1(t_0 + h)| < \varepsilon/4D,$$
$$K/(|h|q^{1/2}) < \varepsilon/4.$$

For $q > Q$ we then have

$$\frac{|\ell_q^2\Phi(qy_q^1(t_0))|}{q} < \varepsilon/4 + D\{|\ell_q^1(t_0 + h) - \ell_0^1(t_0 + h)| + |\ell_0^1(t_0 + h) - \ell_0^1(t_0)| + |\ell_0^1(t_0) - \ell_q^1(t_0)|\} < \varepsilon,$$

and the lemma follows.

Now consider equation (4.2b). According to this result the function $\mu_q(t)$ defined by

$$\text{(4.4e)} \quad \mu_q(t) = \inf_{u \in U} \{F(t, x_q(t), u, \lambda_q(t)) + b(t, x_q(t), u) \cdot \ell_q^1(t) + \ell_q^2\Phi(q \cdot y_q^1(t))/q\}$$

satisfies the integral equation

$$\text{(4.4f)} \quad \mu_q(t) = C_q + \int_\alpha^t \mathfrak{M}_q[F_t(\gamma, x_q(\gamma), u, \lambda_q(\gamma)) + b_t(\gamma, x_q(\gamma), u) \cdot \ell_q^1(\gamma); \gamma]\, d\gamma$$

for each q in Γ and t in T. The function $F(t, x, u, \lambda) + b(t, x, u) \cdot \ell$ is continuous in all of its variables, so whenever $\ell_0^1(t)$ is continuous, Lemma 4.4c shows that the right-hand side of (4.4e) converges to

$$\inf_{u \in U} \{F(t, x_0(t), u, \lambda_0(t)) + b(t, x_0(t), u) \cdot \ell_0^1(t)\} + 0$$

as q increases in Γ. Moreover, if we choose t_0 to be some point in $[a_0, b_0]$ such that $\ell_0^1(t)$ is continuous at $t = t_0$, and rewrite the equation (4.4f) in the form

$$\mu_q(t) = \mu_q(t_0) + \int_{t_0}^t \mathfrak{M}_q[F_t(\gamma, x_q(\gamma), u, \lambda_q(\gamma)) + b_t(\gamma, x_q(\gamma), u)\ell_q^1(\gamma); \gamma]\, d\gamma,$$

an argument like that which established (4.4b) shows that Γ can be chosen so that the functions $\mu_q(t)$ converge to a function $\mu_0(t)$ which satisfies the equation

$$\text{(4.4g)} \quad \mu_0(t) = \mu_0(\alpha) + \int_\alpha^t \mathfrak{M}_0[F_t(\gamma, x_0(\gamma), u, \lambda_0(\gamma)) + b_t(\gamma, x_0(\gamma), u) \cdot \ell_0^1(\gamma); \gamma]\, d\gamma.$$

From the remarks above it is evident that

$$\text{(4.4h)} \quad \mu_0(t) = \inf_{u \in U} \{F(t, x_0(t), u, \lambda_0(t)) + b(t, x_0(t), u)\ell_0^1(t)\}$$

for almost all t in $[a_0, b_0]$. In particular, equality must hold whenever $\ell_0^1(t)$ is continuous. This establishes conclusion (iii) of Theorem 3.2a.

Choose a point t_0 in $[a_0, b_0]$ at which $\ell_q^2 \Phi(q \cdot y_q^1(t))/q$ approaches zero and $\ell_0^1(t)$ is continuous. These conditions both hold for almost all t_0 in $[a_0, b_0]$ since $\ell_0^1(t)$ is bounded and monotone, and therefore is continuous almost everywhere. Now let u_0 be a point in the support of $\mathfrak{M}_0$ at t_0, and suppose that (t_j, u_j) is a sequence of pairs, with u_j in the support of $\mathfrak{M}_j$ at t_j, such that (t_j, u_j) converges to (t_0, u_0) as j increases. Further suppose that $\ell_q^2 \cdot \Phi(q \cdot y_q^1(t_j))/q$ approaches zero for each $t_j, j = 1, 2, \cdots$, as q increases in Γ. Such a sequence of pairs exists according to (iii) of Theorem 2.3d. Now we use (4.2c) to obtain

$$\mu_0(t_0) = \lim_{j \to \infty} \lim_{q \to \infty} (\mu_q(t_j))$$

$$= \lim_{j \to \infty} \lim_{q \to \infty} \{F(t_j, x_q(t_j), u_j, \lambda_q(t_j)) + \ell_q^1(t_j) \cdot b(t_j, x_q(t_j), u_j) + \ell_q^2 \cdot \Phi(q \cdot y_q^1(t_j))/q\} \tag{4.4i}$$

$$= F(t_0, x_0(t_0), u_0, \lambda_0(t_0)) + \ell_0^1(t_0) b(t_0, x_0(t_0), u_0) + 0.$$

This is conclusion (iv) of Theorem 3.2a.

4.5. Transversality conditions: Condition (v). Suppose that η_0 is some point in E and let E' be the set of inward directions to E at η_0. By definition, E' is the set of vectors of the form $L'(0; h)$ defined in (3.1d). The differential $L'(0; z)$, z in R^{2n+2}, is a linear map which splits R^{2n+2} into the kernel, $\ker(L'(0; z))$, and the orthogonal complement of this kernel, which we shall denote R^+. The space R^+ is a Euclidean space of dimension $v \leqq n$, and the restriction of the original map L to R^+ is one-to-one within some closed neighborhood N about 0 in R^{2n+2}. Moreover, L has a continuously differentiable inverse on the set $L(N \cap R^+)$. Now consider the set $Q = N \cap K \cap R^+$. The image $L(Q)$ is a subset of E which contains η_0 and there exists a continuously differentiable map L^{-1} from $L(Q)$ back onto Q. The set $R^+ \cap K$ is a convex cone in R^{2n+2} and for each η in $L(Q)$ we define the set E'_η to be the set of vectors of the form $L'(L^{-1}(\eta); h)$, where h is in the cone R^+K. We adjust this choice so that if $L^{-1}(\eta)$ is on the boundary of N, then $E'_\eta = 0$.

We have gone through this rather complicated construction in order to choose a subset $L(Q)$ of E containing η_0, and inward directions E'_η defined for η in $L(Q)$ such that if η approaches η_0 in $L(Q)$, then E'_η approaches E'_{η_0}. More precisely, for any vector $L'(0, h)$ in E'_{η_0} we have

$$L'(0; h) = \lim_{\eta \to \eta_0} L'(L^{-1}(\eta); h), \qquad \eta \in L(Q). \tag{4.5a}$$

This property will be necessary in order to derive conclusion (v) in Theorem 3.2a as a limiting case of (4.2d).

Suppose that C_0 is the optimal curve obtained in § 4.3, and let $\eta(C_0)$ denote its endpoint in E. Using the construction of the previous two paragraphs we set $\eta(C_0) = \eta_0$ and choose a set $L(Q)$ which contains $\eta(C_0)$, and a collection of inward directions E'_η defined for η in $L(Q)$ which satisfy (4.5a). Now let us replace the set E in our original problem with the set $L(Q)$ and the sets E' with the sets E'_η as we have just defined them. Furthermore, let us define a functional J' by

$$J'(C) = J(C) + \|\eta(C) - \eta(C_0)\|^2,$$

where C is any admissible curve for this new problem, and $\eta(C)$ denotes the endpoint of C in $L(Q)$. Here $\|\cdot\|$ denotes the Euclidean norm on R^{2n+2}. This new problem satisfies the hypotheses of Theorem 3.2a and we conclude that there must exist an optimal curve C'_0 which avoids S' and which satisfies conclusions (i) through (iv) which we have already established. Suppose that C'_0 is defined by

$$C'_0 = \lim_{q\to\infty} C'_q, \qquad q \text{ in } \Gamma',$$

where C'_q and Γ' play the roles of C_q and Γ in § 4.1 through § 4.4, and suppose the functions $\lambda_q^{0\prime}, \lambda_q^{1\prime}, \cdots, \lambda_q^{n\prime}, \ell_q^{1\prime}$ similarly correspond to $\lambda_q^0, \lambda_q^1, \cdots, \lambda_q^n, \ell_q^1$. Since the function $e(t_0, x_0, t_1, x_1)$ does not appear in conditions (i) through (iv), no expression involving $\|\eta(C) - \eta(C_0)\|^2$ appears either, and these conditions hold unchanged for the curve C'_0. Furthermore, since C_0 is admissible for the new problem, we must have $J'(C'_0) = J(C'_0) = J(C_0)$, and consequently,

$$\begin{aligned} 0 &= \|\eta(C'_0) - \eta(C_0)\| \\ &= \lim_{q\to\infty} \|\eta(C'_q) - \eta(C_0)\|, \qquad q \text{ in } \Gamma'. \end{aligned}$$

This means that $\eta(C'_q)$ approaches $\eta(C_0)$ as q increases in Γ', and hence C'_0 also has endpoint $\eta(C_0)$. Thus expressions involving partial derivatives of $\|\eta(C) - \eta(C_0)\|^2$ go to zero in the appropriate versions of (4.2d). Moreover if $L'(L^{-1}(\eta(C_0)); h)$ is any vector in $E'_{\eta(C_0)}$, then

$$L'(L^{-1}(\eta(C_0)); h) = \lim_{q\to\infty} L'(L^{-1}(\eta(C'_q)); h), \qquad q \text{ in } \Gamma',$$

and $L'(L^{-1}(\eta(C'_q)); h)$ is in $E'_\eta(C'_q)$. Condition (v) follows by taking limits in (4.2d) after setting w_1^2 equal to zero for each curve C'_q.

5. Proof of Theorems 3.3a and 3.3b.

5.1. Proof of Theorem 3.3a. We shall prove this theorem by showing that the function $\zeta(t)$ defined by

$$\zeta(t) = \left\{ \sum_{i=0}^{n} (\lambda^i(t))^2 \right\} + (\ell(t))^2$$

cannot equal zero anywhere on the interval $[a, b]$ unless it is identically zero on $(a, b]$. Assume that $\zeta(t)$ does equal zero somewhere in $[a, b]$. We shall show this implies that $\zeta(t) \equiv 0$ on $(a, b]$. The first conclusion from our assumption is that the constant λ^0 must be zero. Now let t^* be the infimum of the t values for which $\zeta(t)$ is zero, and let s^* be the infimum of the points at which $\ell(t)$ is zero. Evidently $s^* \leqq t^*$, and since each of the functions $\lambda^i(t)$, $i = 0, 1, 2, \cdots, n$, is continuous, the function $\zeta(t)$ is continuous whenever $\ell(t)$ is continuous. Next we use Lemma 4.2g, which extends easily to apply here, to show

$$\begin{aligned} \sum_{i=0}^{n} (\lambda^i(t))^2 &\leqq e^{K_1|t-t^*|} \sum_{i=0}^{n} (\lambda^i(t^*))^2 + K_2 \left| \int_t^{s^*} (\ell(\gamma))^2 \, d\gamma \right| + K_2(\lambda^0)^2 |t - t^*| \\ &= K_2 \left| \int_t^{s^*} (\ell(\gamma))^2 \, d\gamma \right|, \end{aligned} \tag{5.1a}$$

and since $\ell(t)$ is monotone decreasing, we see that $\zeta(t)$ is identically zero for $t > s^*$.

Now consider two cases. Suppose first that $s^* = a$. In this case (5.1a) shows that the functions $\lambda^0, \lambda^1(t), \cdots, \lambda^n(t)$ must all be zero in the entire interval $[a, b]$, and furthermore, $\ell(t)$ must be zero everywhere in $(a, b]$. Thus $s^* = t^* = a$. We remark that if $p(x(a)) < 0$, then s^* cannot equal a. This is true since $\zeta(t)$ must be continuous at a in this case, and by the remarks following (4.4a), if $p(x(a))$ is negative, then the vector $(\lambda^0, \lambda^1(a), \cdots, \lambda^n(a), \ell(a))$ is not zero. It follows that there must be an interval about a on which $\zeta(t)$ is not zero. This is impossible since $t^* = a$. Suppose on the other hand that $s^* > a$. Let K be a bound for the functions $f^i(t, x(t), u), i = 0, 1, \cdots, n$, and observe that since $\lambda^0 = 0$, we may use (5.1a) to show that for $t < t^*$,

$$
\begin{aligned}
|F(t, x(t), u, \lambda(t))| &= \left| \sum_{i=0}^{n} f^i(t, x(t), u) \cdot \lambda^i(t) \right| \\
&\leqq K \cdot \sum_{i=1}^{n} |\lambda^i(t)| + K \cdot |\lambda^0| \\
&\leqq K \cdot n \cdot \left[\sum_{i=1}^{n} (\lambda^i(t))^2 \right]^{1/2} + 0 \\
&\leqq K \cdot n \cdot \left[K_2 \cdot \int_t^{t^*} (\ell(\gamma))^2 \, d\gamma \right]^{1/2} \\
&\leqq K_2 K \cdot n \cdot [(\ell(t))^2 \cdot |t^* - t|]^{1/2} = K_2 K n \cdot \ell(t) \cdot [|t^* - t|]^{1/2}.
\end{aligned}
$$

Conditions (iii) and (iv) of Theorem 3.2a taken together show that for u_0 in the support of $\mathfrak{M}$ at t, and almost all t in $[a, b]$,

$$
\begin{aligned}
&F(t, x(t), u_0, \lambda(t)) + \ell(t) \cdot b(t, x(t), u_0) \\
&\quad = \inf_{u \in U} [F(t, x(t), u, \lambda(t)) + \ell(t) \cdot b(t, x(t), u)] \\
&\quad \leqq nKK_2 \ell(t) \cdot [|t^* - t|]^{1/2} + \inf_{u \in U} [\ell(t) \cdot b(t, x_0(t), u)].
\end{aligned}
$$

Now since s^* is the infimum of the t values for which $\ell(t) = 0$, and $\ell(t)$ is constant on any interval on which $p(x(t))$ is negative, we conclude that $p(x(s^*)) = 0$. It follows from our assumption that $\inf_{u \in U} b(s^*, x(s^*), u) \leqq -v < 0$ and hence there is an interval about s^* on which $\inf_{u \in U} b(t, x(t), u) \leqq -v/2$. For these values of t we have

$$
\begin{aligned}
\ell(t) \cdot b(t, x(t), u_0) &\leqq nKK_2 \cdot \ell(t) \cdot [|t^* - t|]^{1/2} + \inf_{u \in U} [\ell(t) \cdot b(t, x_0(t), u)] \\
&\quad - F(t, x(t), u_0, \lambda(t)) \\
&\leqq nKK_2 \cdot \ell(t) \cdot [|t^* - t|]^{1/2} - \ell(t)v/2 + nKK_2 \cdot \ell(t) \cdot [|t - t^*|^{1/2}] \\
&= 2nKK_2 \cdot \ell(t) \cdot [|t^* - t|]^{1/2} - (v/2) \cdot \ell(t) \quad \text{a.e.}
\end{aligned}
$$

Dividing through by $\ell(t)$, t less than s^*, gives

$$
\begin{aligned}
b(t, x(t), u_0) &\leqq 2nKK_2 \cdot [|t^* - t|]^{1/2} - v/2 \\
&\leqq -v/4 < 0
\end{aligned}
$$

for almost all t near s^* on the left, and u_0 in the support of $\mathfrak{M}$ at t. But since

$$p(x(t)) = p(x(a)) + \int_a^t \mathfrak{M}[b(\gamma, x_0(\gamma), u); \gamma]\, d\gamma,$$

$p(x(t))$ must be a decreasing function as $t \to s^*$ from below. This is impossible since $p(x(s^*))$ is a maximum value for $p(x(t))$ on $[\alpha, \beta]$. This contradiction completes our proof.

5.2. Proof of Theorem 3.3b. Suppose that $C_0 = (\mathfrak{M}_0, x_0(t), [a_0, b_0])$ is some particular S-optimal curve. We wish to prove Theorem 3.3b by devising a problem in which C_0 is the only possible optimal curve, and then apply Theorems 3.2a and 3.3a to this problem. We need a penalty function which penalizes any deviation of a curve C from the trajectory $x_0(t)$. This trajectory is not sufficiently differentiable as a function of t, so we must approximate $x_0(t)$ by continuously differentiable functions $h_q(t)$. We proceed by letting $h_1(t), h_2(t), \cdots$ be a sequence of twice continuously differentiable functions with uniformly bounded first derivatives which converges uniformly to the continuous function $x_0(t)$ on the interval $[\alpha, \beta]$. Such a sequence can be constructed using integral means as suggested in Graves [5], or by using the fact that x_0 is Lipschitz and approximating it with polygonal curves whose corners have been "rounded off." Next construct the functional J_q on the collection of S-admissible curves by defining

$$J_q(C) = J(C) + \sum_{i=1}^{n} \int_\alpha^\beta [(x^i(\gamma) - h_q^i(\gamma))^2; \gamma]\, d\gamma + \|\eta(C) - \eta(C_0)\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm in R^{2n+2}. Finally we repeat the construction of § 4.5 to find a set $L(Q)$ and sets E'_η, $\eta \in L(Q)$, such that for η near $\eta(C_0)$, $E'_\eta \to E'_{\eta(C_0)}$ as $\eta \to \eta(C_0)$ in $L(Q)$. We may now apply Theorem 3.2a to find a curve C_q satisfying conclusions (i) through (v) of that theorem, such that C_q optimizes J_q over the class of S-admissible curves whose endpoints lie in the set $L(Q)$. The curves C_q all have trajectories which lie in a single closed and bounded subset of R^{n+1}, so by Theorem 2.3d there exists a sequence Γ of integers and a generalized curve $C_0^* = (\mathfrak{M}_0^*, x_0^*, [a_0^*, b_0^*])$ such that $C_0^* = \lim_{q \to \infty} C_q$, for q in Γ. Now for any q in Γ,

$$\begin{aligned} J(C_q) &\leqq J_q(C_q) \leqq J_q(C_0) \\ &= J(C_0) + \sum_{i=1}^{n} \int_\alpha^\beta \mathfrak{M}_0[(x_0^i(\gamma) - h_q^i(\gamma))^2; \gamma]\, d\gamma + \|\eta(C_0) - \eta(C_0)\|^2, \end{aligned}$$

and this last expression goes to $J(C_0)$ as q increases in Γ. Since J is continuous this means that $J(C_0^*) \leqq J(C_0)$, and hence the two values are equal. It follows that as q increases in Γ, $J_q(C_q)$ approaches $J(C_0)$ so

$$\lim_{q \to \infty} \int_\alpha^\beta \mathfrak{M}_q[(x_q^i(\gamma) - h_q^i(\gamma))^2; \gamma]\, d\gamma = 0, \qquad q \text{ in } \Gamma,$$

and

$$\lim_{q \to \infty} \|\eta(C_q) - \eta(C_0)\|^2 = 0.$$

This means that C_0^* has the same trajectory and endpoints as C_0. It remains to show that Γ can be chosen so that the conclusions of Theorem 3.2a hold.

Since $\sum_{i=0}^{n} \lambda_q^i(\alpha)^2 + \ell_q(\alpha)^2 \leqq 1$, we may use Helley's selection theorem to show that λ_q^0 and ℓ_q converge to functions λ_*^0 and ℓ_* satisfying (i). For any C_q, formula (ii) has the form

$$\lambda_q^i(t) = \lambda_q^i(a_q) - \int_{a_q}^{t} \mathfrak{M}_q[F_{x^i}(\gamma, x_q(\gamma), u, \lambda_q(\gamma)) + \ell(\gamma) b_{x^i}(\gamma, x_q(\gamma), u); \gamma]\, d\gamma$$

$$- \int_{a_q}^{t} 2(x_q^i(\gamma) - h_q^i(\gamma))\, d\gamma$$

and an argument like that of (4.4b) along with the uniform convergence of the functions h_q to x_0 establishes condition (ii). Formula (iii) becomes

$$\mu_q(t) = \inf_{u \in U} \{F(t, x_q(t), u, \lambda_q(t)) + \ell_q(t) b(t, x_q(t), u) + \lambda_q^0 \sum_{i=1}^{n} (x_q^i(t) - h_q^i(t))^2\}$$

$$= c_q + \int_{a_q}^{t} \mathfrak{M}_q[F_t(\gamma, x_q(\gamma), u, \lambda_q(\gamma)) + \ell_q(\gamma) b_t(\gamma, x_q(\gamma), u); \gamma]\, d\gamma$$

$$+ \sum_{i=1}^{n} \int_{a_q}^{t} [2(x_q^i(\gamma) - h_q^i(\gamma))][h_q^i(\gamma)]\, d\gamma$$

and the desired formula follows directly at almost all points in $[a_0^*, b_0^*]$. The second expression above is continuous in the limit wherever ℓ is continuous, so formula (iii) is verified.

Formula (iv) follows by letting A' be the union of the sets A_q, $q \in \Gamma$, described in condition (iv) for each C_q, and letting (t_q, u_q) be a pair with t_q in $\{[\alpha - \beta] - A'\}$ and u_q in the support of $\mathfrak{M}_q$ at t_q. We now apply Theorem 2.3d as for (4.4h). Now condition (v) follows in a straightforward fashion since $E'_{\eta(C_q)} \to E'_{\eta(C_0)}$ as q increases.

6. An example. Let us consider the problem of finding a curve in n-space which has minimum arc length; extends between distinct points π_0, and π_1; and remains outside an infinite open cylinder of radius one about the origin. For simplicity we shall suppose that π_0 and π_1 are interior to S.

It is clear that there is a curve between π_0 and π_1 which remains clear of the forbidden cylinder, and we let b be the arc length of one such curve. Using arc length as a curve parameter, we formulate the problem as follows. Let T be the interval $[0, 2b]$, let B be the ball of radius $2b$ about π_0 in n-space, and let U be the unit sphere in n-space. Then for $t \in T$, $u \in U$, $x \in B$ define

$$f^0(t, x, u) = 1, \qquad f^i(t, x, u) = u^i, \qquad i = 1, 2, \cdots, n, \tag{6.1}$$

where

$$\sum_{i=1}^{n} (u^i)^2 = 1.$$

Let p be defined by

$$p(x) = 1 - \sum_{k=1}^{m} (x^k)^2, \tag{6.2}$$

where $m \leqq n$. Clearly p can be adjusted so that S is in B, but since this will not alter our results, we shall not make the adjustment. Now define the set E to be $\{0\} \times \{\pi_0\} \times T \times \{\pi_1\}$, and notice that for $0 < t < 2b$, the set E' of inward directions at $(0, \pi_0, t, \pi_1)$ is the set $\{0\} \times \{0\} \times R^1 \times \{0\}$. In this setting, the problem we posed is the problem of finding a curve C_0 and an interval $[0, b_0]$ such that the value

$$J(C_0) = \int_0^{b_0} 1 \, d\gamma = b_0 \tag{6.3}$$

is the smallest attained by any admissible curve whose trajectory remains outside the forbidden set $\{x : p(x) > 0\} = S'$.

The problem stated here satisfies the hypotheses of Theorems 3.2a, 3.3a and 3.3b, and the point π_0 is not on the boundary of the forbidden region, so there exists a minimizing generalized curve $C_0 = (\mathfrak{M}_0, x_0, [0, b_0])$ along with corresponding functions $\lambda^0, \lambda^1, \cdots, \lambda^n, \ell$ which satisfy the conclusions of Theorem 3.2a. Moreover these functions are not all zero at any point in the interval $[0, b_0]$. We have from Theorem 3.2a(ii),

$$\lambda^i(t) = \begin{cases} \lambda^i(0) + \displaystyle\int_0^t 2\ell(\gamma)\mathfrak{M}_0[u^i; \gamma] \, d\gamma, & i = 1, \cdots, m, \\ \lambda^i(0), & i = m+1, \cdots, n. \end{cases} \tag{6.4}$$

Furthermore, the differential equation in (iii) shows that $\mu(t)$ is constant, and since condition (v) reduces to the inequality $\mu(b_0)\gamma_1 \geqq 0$ for all real γ_1, we conclude that μ is identically zero. If we adopt the notation

$$\sum_{i=1}^{n} x_1^i \cdot x_2^i = \langle x_1, x_2 \rangle$$

for the inner product between vectors x_1 and x_2 in Euclidean R^n, then we can combine the results in (iii) and (iv) to obtain

$$\begin{aligned} 0 &= \lambda^0 + \langle \lambda(t) + \ell(t) p_x(x_0(t)), u_0 \rangle \\ &= \inf_{u \in U} \{\lambda^0 + \langle \lambda(t) + \ell(t) p_x(x_0(t)), u \rangle\} \end{aligned} \tag{6.5}$$

for almost all t in $[0, b_0]$, and for u_0 in the support of $\mathfrak{M}_0$ at t. This equation has a unique solution $u_0(t)$ unless $\lambda(t) + \ell(t) p_x(x_0(t)) = 0$. The last alternative implies that $\lambda^0 = 0$, and since $\ell(b_0) = 0$, the last term in (6.5) would be negative at t_0. We conclude that there is a unique point $u_0(t)$ in the support of $\mathfrak{M}_0$ at t for almost all t in $[0, b_0]$, so we may write C_0 as an ordinary curve with control function

$$u_0(t) = -\frac{\lambda(t) + \ell(t) p_x(x_0(t))}{\| \lambda(t) + \ell(t) p_x(x_0(t)) \|}, \tag{6.6}$$

where $\| \cdot \|$ denotes the Euclidean norm in R^n. Substitution of (6.6) into (6.5) shows that

$$\lambda^0 = \| \lambda(t) + \ell(t) p_x(x_0(t)) \| \neq 0 \tag{6.7}$$

and

$$u_0(t) = -\frac{\lambda(t) + \ell(t)p_x(x_0(t))}{\lambda^0}. \tag{6.8}$$

Now let us try to evaluate $\ell(t)$. If we square equation (6.7) and use inner product notation, we get a quadratic equation in ℓ which has the solution

$$\begin{aligned}\ell(t) = &\frac{-\langle p_x(x_0(t)), \lambda(t)\rangle}{\langle p_x(x_0(t)), p_x(x_0(t))\rangle} \\ &\pm \frac{\{\langle p_x(x_0(t)), \lambda(t)\rangle^2 - \langle p_x(x_0(t)), p_x(x_0(t))\rangle[\langle \lambda(t), \lambda(t)\rangle - \lambda^0]^2\}^{1/2}}{\langle p_x(x_0(t)), p_x(x_0(t))\rangle}.\end{aligned} \tag{6.9}$$

On the other hand,

$$\begin{aligned}b(t, x_0(t), u_0(t)) &= \langle p_x(x_0(t)), u_0(t)\rangle \\ &= \frac{-\langle p_x(x_0(t)), \lambda(t)\rangle}{\lambda^0} - \ell(t)\frac{\langle p_x(x_0(t)), p_x(x_0(t))\rangle}{\lambda^0}\end{aligned} \tag{6.10}$$

so that solving for $\ell(t)$ gives

$$\ell(t) = \frac{-\langle p_x(x_0(t)), \lambda(t)\rangle - \lambda^0 \cdot b(t, x_0(t), u_0(t))}{\langle p_x(x_0(t)), p_x(x_0(t))\rangle}. \tag{6.11}$$

Since zero is a relative maximum for $p(x_0(t))$ the sign in (6.10) must change from $\geqq 0$ to $\leqq 0$ as t passes any area in which $p(x_0(t)) = 0$. This means that the sign of the square root in (6.9) must go from negative to positive, and since $\ell(t)$ is monotonely decreasing, this means that $b(t, x_0(t), u_0(t)) = 0$ where the sign change occurs. In fact, since $b(t, x_0(t), u_0(t))$ is the derivative of $p(x_0(t))$, it must be zero throughout any interval on which $p(x_0(t)) = 0$ identically, and continuity of the expression under the radical in (6.9) assures that $\ell(t)$ is continuous at the ends of such an interval. Therefore,

$$\ell(t) = -\frac{\langle p_x(x_0(t)), \lambda(t)\rangle}{\langle p_x(x_0(t)), p_x(x_0(t))\rangle} \tag{6.12}$$

whenever $p(x_0(t)) = 0$. Theorem 3.2a states that $\ell(t)$ is constant and hence continuous wherever $p(x_0(t)) < 0$, and the discussion above shows that $\ell(t)$ is continuous elsewhere. It follows that if $p(x_0(t)) = 0$ at an isolated point $t = t_0$, then $\ell(t)$ remains constant at t_0. Now suppose that $p(x_0(t)) = 0$ on some interval $a_0 \leqq t \leqq a_1$. The definition of $p(x)$ shows that on this interval we have

$$\langle p_x(x_0(t)), p_x(x_0(t))\rangle \equiv \sum_{i=1}^{m} (-2x_0^i(t))^2 = 4.$$

Differentiating (6.12) and using (6.4) and the fact that $b = 0$ on this interval, we have

$$\begin{aligned}\frac{d}{dt}\ell &= -\frac{1}{4}\frac{d}{dt}\langle p_x(x_0(t)), \lambda(t)\rangle \\ &= -\frac{1}{4}\left[\sum_{i=1}^{m}\frac{d}{dt}(-2x_0^i(t))\cdot\lambda^i(t) + \sum_{i=1}^{m} -2x_0^i(t)\frac{d}{dt}\lambda^i(t)\right] \\ &= \frac{1}{2}\sum_{i=1}^{m} u_0^i(t)\cdot\lambda^i(t) + \frac{1}{2}\sum_{i=1}^{m} x_0^i(t)\cdot 2\ell(t)\cdot u_0^i(t) \\ &= \frac{1}{2}\sum_{i=1}^{m} u_0^i(t)\cdot[-\lambda^0 u_0^i(t) - \ell 2x_0^i(t)] + 0 = \frac{-\lambda^0}{2}\cdot\sum_{i=1}^{m} u_0^i(t)^2\end{aligned}$$

for $a_0 \leqq t \leqq a_1$.

Let us return to our previous expression for $u_0(t)$ given in (6.8). This function is continuous and piecewise differentiable with derivatives given by

$$\begin{aligned}\frac{d}{dt}u_0^i(t) &= \frac{-2\ell(t)u_0^i(t) + \ell(t)2u_0^i(t) + [d\ell(t)/dt]2x_0^i(t)}{\lambda^0} \\ &= \begin{cases} -\left[\sum_{j=1}^{m}[u_0^j(t)]^2\right]\cdot x_0^i(t) & \text{if } p(x_0(t)) = 0, \quad i = 1, 2, \cdots, m, \\ 0 & \text{if } p(x_0(t)) < 0, \quad i = 1, 2, \cdots, m, \end{cases}\end{aligned}$$

and

$$\frac{du_0^i(t)}{dt} = 0 \quad \text{if } 0 \leqq t \leqq b_0, \quad i = m+1, \cdots, n.$$

The definition of the set U required that $\sum_{i=1}^{n}(u^i)^2 = 1$, and since the components $u_0^i(t)$ are constant for $i = m+1, m+2, \cdots, n$, this means that

$$\sum_{i=1}^{m}(u_0^i(t))^2 = 1 - \sum_{i=m+1}^{n}(u_0^i)^2 = K \geqq 0$$

for some constant K.

Summarizing the previous analysis, we have shown that any optimal curve C_0 for the problem we posed must be an ordinary curve whose trajectory $x_0(t)$ is a straight line unless it touches the cylinder $S' = \{x : p(x) > 0\}$. The optimal trajectory must meet the cylinder tangentially and satisfy the differential equations

$$d^2x^i/dt^2 = -Kx^i, \qquad i = 1, 2, 3, \cdots, m,$$

with initial conditions provided by the continuity of u and x. These are the equations for simple harmonic motion. Further,

$$d^2x^i/dt^2 = 0 \quad \text{for } 0 \leqq t \leqq b_0, \qquad i = m+1, m+2, \cdots, n,$$

and these coordinates are linear functions of t.

Acknowledgment. The author is indebted to Professor E. J. McShane for his encouragement and for his valuable comments and corrections to early versions of this paper.

REFERENCES

[1] E. J. McShane, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.

[2] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[3] J. Warga, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432–455.

[4] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Soc. Sci. Lettres de Varsovie, Classe III, 30 (1937), pp. 212–234.

[5] L. M. Graves, *Some general approximation theorems*, Ann. of Math., 42 (1941), pp. 281–292.

[6] S. S. L. Chang, *Optimal control in bounded phase space*, Automatica, 1 (1962), pp. 55–67.

[7] L. W. Neustadt, *An abstract variational theory with applications to a broad class of optimization problems. II: Applications*, this Journal, 5 (1967), pp. 90–138.

[8] ———, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

[9] A. Ya. Dubovitskii and A. A. Milyutin, *Extremum problems in the presence of constraints*, Zh. Vychisl. Mat. i Mat. Fiz., 5 (1965), pp. 395–453; English transl., *Problems on the extremum in the presence of limitations*, SLA Translations Center, no. TT-66-10916.

[10] F. A. Valentine, *The problem of Lagrange with differential inequalities as side conditions*, Contributions to the Calculus of Variations 1933–1937, University of Chicago Press, Chicago, 1937, pp. 407–448.

[11] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

THE USE OF STATE FEEDBACK FOR EXACT MODEL MATCHING*

W. A. WOLOVICH†

Abstract. A concise feedback "characterization" of a linear multivariable system, expressed in terms of a frequency domain "factorization" of the transfer matrix, is defined and employed to completely and concisely represent a large class of systems compensated by state feedback. This "characterization" is then employed to answer the question of "exact model matching."

1. Introduction. In this paper, we present a complete and concise frequency domain "characterization" of linear, state form multivariable systems subjected to linear state variable feedback (l.s.v.f.) and then utilize the "characterization" to provide, for the first time, a solution to a fundamental question pertaining to the synthesis of linear multivariable systems, namely the question of "exact model matching."

In § 2, we discuss the significance of the characterization from the point of view of system synthesis. In particular, we precisely formulate the question of "exact model matching" and then compare the formulation to some recent, and closely related, work. In § 3, we employ a transformation to "companion canonical form" for the time domain dynamical equations in order to develop the frequency domain characterization in the scalar case. This serves to motivate the more general multivariable case, which we consider in § 4. Again, we employ a certain "companion form" for the time domain dynamics which enable us to develop the "characterization" in the multivariable case.

In § 5, we employ the "characterization" in order to provide new insight into the synthesis of linear multivariable systems. In particular, we constructively resolve the question of "exact model matching," as posed in § 2, and then conclude with a discussion of possible extensions in § 6.

2. The problem. We begin by considering the linear, time-invariant dynamical system, $\{A, B, C, D\}$; i.e., the system

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{1a}$$

$$y(t) = Cx(t) + Du(t), \tag{1b}$$

where $x(t)$ is an n-vector called the state, $u(t)$ is an m-vector called the input, $y(t)$ is a p-vector called the output, and A, B, C and D are constant real $n \times n$, $n \times m$, $p \times n$ and $p \times m$ matrices respectively. Without much loss of generality, we assume that the pair $\{A, B\}$ is completely controllable, and that B is of full rank $m \leqq n$. We next consider the linear state variable feedback (l.s.v.f.) control law:

$$u(t) = Fx(t) + Gw(t), \tag{2}$$

where $w(t)$ is an m-vector called the external input, and the feedback pair, $\{F, G\}$, are constant real $m \times n$ and $m \times m$ matrices respectively, with G assumed nonsingular.

* Received by the editors April 8, 1971, and in revised form October 31, 1971.

† Division of Engineering, Brown University, Providence, Rhode Island 02912. This research was supported in part by the National Science Foundation—Engineering under Grant GK-24485, and in part by the Air Force Office of Scientific Research under Grant AF-AFOSR 693-67D.

This latter assumption ($|G| \neq 0$) is necessary in order to insure the linear independence of the (m) external inputs, $w(t)$. Substituting (2) for $u(t)$ in (1) results in the closed loop system:

$$\dot{x}(t) = (A + BF)x(t) + BGw(t), \tag{3a}$$

$$y(t) = (C + DF)x(t) + DGw(t). \tag{3b}$$

A fundamental question which now arises is the following: Is it possible completely and concisely to characterize the entire class of closed loop systems of the form (3) for all possible feedback pairs $\{F, G\}$? An affirmative answer to this question would provide a means of resolving a number of practical synthesis questions such as: Does there exist a feedback pair $\{F, G\}$ such that the quadruple $\{A + BF, BG, C + DF, DG\}$ is "similar" to a given quadruple $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$; i.e., such that $P(A + BF)P^{-1} = \tilde{A}$, $PBG = \tilde{B}$, $(C + DF)P^{-1} = \tilde{C}$, and $DG = \tilde{D}$ for some nonsingular matrix P? More generally, does there exist a feedback pair $\{F, G\}$ such that the transfer matrix of the closed loop system, namely $(C + DF) \cdot (sI - A - BF)^{-1}BG + DG$, is identical to the transfer matrix $T_m(s)$ of a "model" system? This latter question, which will be constructively resolved in § 5, will be called *the question of exact model matching via l.s.v.f.*, and has been shown to be of practical importance in the design of "model following" control systems [1], [2]. More specifically, Erzberger [1] appears to have been first to formulate and solve a rather restrictive version of the question of exact model matching via l.s.v.f. just posed. In particular, the order of his model ($\dot{z}(t) = Lz(t)$) is restricted to be equal to the dimension of the output vector y. He then obtains only sufficient conditions under which "perfect model following" is possible. More recently, Winsor and Roy [2] have included Erzberger's work in a somewhat different, although not original, treatment of "exact model following," one which involves formulating the question as one in linear optimal control. The solution to the problem, thus formulated, involves actual simulation of the model, as part of the control scheme, as well as feedforward compensation (from the model) in combination with partial state feedback.

The results which we shall present in the remainder of this paper employ a frequency domain (transfer matrix), rather than a time domain model as employed by Erzberger and Windsor and Roy. Nevertheless, the same basic question of model matching is considered here as in the closely related studies noted above.

3. The scalar case. Before discussing the multivariable case, we first consider the rather simple, single input/output (scalar) case in order to motivate the more general result. In particular, consider the system:

$$\dot{x}(t) = Ax(t) + bu(t), \tag{4a}$$

$$y(t) = cx(t) + du(t), \tag{4b}$$

where b and c are n-dimensional column and row vectors respectively, and d is a scalar. We again assume that the pair $\{A, b\}$ is completely controllable, which enables us to find an $n \times n$ nonsingular matrix Q which transforms the scalar system (4) to "controllable companion form"; i.e., if we let $z(t) = Qx(t)$, where Q

is obtained from the inverse of the controllability matrix $K = [b, Ab, \cdots, A^{n-1}b]$, then [3], [4]:

$$\dot{z}(t) = \hat{A}z(t) + \hat{b}u(t), \tag{5a}$$

$$y(t) = \hat{c}z(t) + du(t), \tag{5b}$$

where

$$QAQ^{-1} = \hat{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & \cdot & & \\ & & & \cdot & \\ -\hat{a}_0 & -\hat{a}_1 & \cdots & & -\hat{a}_{n-1} \end{bmatrix},$$

and $(Qb)^T = \hat{b}^T = [0, 0, \cdots, 0, 1]$, and $\hat{c} = cQ^{-1}$. Furthermore, $|sI - A| = |sI - \hat{A}| = s^n + \hat{a}_{n-1}s^{n-1} + \cdots + \hat{a}_1 s + \hat{a}_0$. Under the transformation Q, the l.s.v.f. control law (2) becomes:

$$u(t) = \hat{f}z(t) + gw(t), \tag{6}$$

where $fQ^{-1} = \hat{f} = [\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_n]$ and g is a nonzero scalar. Employing this control law in (5) yields the closed loop scalar system:

$$\dot{z}(t) = (\hat{A} + \hat{b}\hat{f})z(t) + \hat{b}gu(t), \tag{7a}$$

$$y(t) = (\hat{c} + d\hat{f})z(t) + dgw(t), \tag{7b}$$

where $\hat{A} + \hat{b}\hat{f}$ is also a companion matrix, whose last (nth) row is equal to the row vector $[\hat{f}_1 - \hat{a}_0, \hat{f}_2 - \hat{a}_1, \cdots, \hat{f}_n - \hat{a}_{n-1}]$, and $\hat{b}g$ is identically zero except for the last (nth) entry, which is equal to g. Furthermore, $\hat{c} + d\hat{f} = [\hat{c}_1 + d\hat{f}_1, \hat{c}_2 + d\hat{f}_2, \cdots, \hat{c}_n + d\hat{f}_n]$. We thus note that only the last (nth) rows of $\hat{A}$ and $\hat{b}$ are affected by the feedback pair $\{\hat{f}, g\}$, and all pertinent information regarding the effect of l.s.v.f. on the pair $\{\hat{A}, \hat{b}\}$ is contained in the nth rows of both. We also note that $\hat{f}$ can be used to specify completely and arbitrarily all (n) entries of the nth row of $\hat{A} + \hat{b}\hat{f}$, which represent the coefficients of the closed loop characteristic equation, or equivalently, the denominator of the closed loop transfer function. Based on these observations, we could now readily develop a complete system "characterization" under l.s.v.f. in the scalar case. As we shall see, however, it will prove beneficial to go one step further, and consider the transfer function of the system.

In particular, once the system has been transformed to controllable companion form, the transfer function can be determined by inspection. For example, using the procedure outlined in [6], we have:

$$\begin{aligned} T_{f,g}(s) &= (c + df)(sI - A - bf)^{-1}bg + dg = (\hat{c} + d\hat{f})(sI - \hat{A} - \hat{b}\hat{f})^{-1}\hat{b}g + dg \\ &= \frac{g[(\hat{c}_1 + d\hat{f}_1) + \cdots + (\hat{c}_n + d\hat{f}_n)s^{n-1}]}{(\hat{a}_0 - \hat{f}_1) + \cdots + (\hat{a}_{n-1} - \hat{f}_n)s^{n-1} + s^n} + dg \qquad (8) \\ &= \frac{(\hat{c}_1 + d\hat{a}_0) + \cdots + (\hat{c}_n + d\hat{a}_{n-1})s^{n-1} + ds^n}{(1/g)[(\hat{a}_0 - \hat{f}_1) + \cdots + (\hat{a}_{n-1} - \hat{f}_n)s^{n-1} + s^n]} = \frac{r(s)}{p_{f,g}(s)}. \end{aligned}$$

We now define $S^T(s) = [1, s, \cdots, s^{n-1}]$, where T denotes the transpose of $S(s)$, and $\hat{A}_n = [-\hat{a}_0, -\hat{a}_1, \cdots, -\hat{a}_{n-1}]$, the last ($n$th) row of $\hat{A}$. In view of these

definitions, the expressions for $r(s)$ and $p_{f,g}(s)$ in (8) can be simplified, i.e., $r(s) = [ds^n + (\hat{c} - d\hat{A}_n)S(s)]$ and $p_{f,g}(s) = (1/g)[s^n - (\hat{A}_n + \hat{f})S(s)]$, which allows us to write $T_{f,g}(s)$ more concisely as:

$$T_{f,g}(s) = r(s)[p_{f,g}(s)]^{-1} = [ds^n + (\hat{c} - d\hat{A}_n)S(s)] \cdot \left[\frac{1}{g}s^n - \frac{1}{g}(\hat{A}_n + fQ^{-1})S(s)\right]^{-1}. \tag{9}$$

In terms of this expression for $T_{f,g}(s)$, it is now clear that l.s.v.f. can be used to assign completely and arbitrarily all (n) poles of the closed loop system, provided the pair $\{A, b\}$ is completely controllable [5]. However, l.s.v.f. does not affect (i) *the numerator, $r(s)$, of the transfer function, or* (ii) *n, the order of the system,*[1] a fact which has long been recognized. Clearly, these two quantities represent a complete and concise *characterization* of the system under l.s.v.f. in the scalar case, since they are both independent as well as sufficient to characterize the entire class of controllable scalar systems of the form (3) for all possible feedback control laws of the form (6), with $g \neq 0$. Our ultimate goal in this paper is to seek an analogous characterization in the multivariable case, which can then be employed to resolve the question of exact model matching via l.s.v.f. as formulated in § 2. These two objectives will be accomplished in the next two sections.

4. The multivariable case. The "characterization" which we shall develop in the multivariable case also depends, as in the scalar case, on a particular "companion form" frequency domain description of the system. More specifically, let us consider the dynamical system (1) with B of full rank $m \leqq n$, and $\{A, B\}$ controllable. Then the $n \times nm$ controllability matrix $K = [B, AB, \cdots, A^{n-1}B]$ has rank n, and it is possible to define a "lexicographic" basis for R_n consisting of the first (n) linearly independent columns of K, possibly reordered [6]. We let L be the matrix whose columns are the elements of the "lexicographic" basis, so that:

$$L = [b_1, Ab_1, \cdots, A^{\sigma_1 - 1}b_1, b_2, \cdots, A^{\sigma_2 - 1}b_2, \cdots, A^{\sigma_m - 1}b_m], \tag{10}$$

where $b_1, \cdots, b_m$ are the ordered columns of B. Setting $d_k = \sum_{i=1}^{k} \sigma_i$, for $k = 1, 2, \cdots, m$, and defining l_k as the d_kth row of L^{-1}, we can show that the matrix Q given by:

$$Q = \begin{bmatrix} l_1 \\ l_1 A \\ \vdots \\ l_1 A^{\sigma_1 - 1} \\ l_2 \\ \vdots \\ l_m A^{\sigma_m - 1} \end{bmatrix} \tag{11}$$

[1] It is, of course, recognized that common "pole-zero factors" may be present in both the numerator and denominator polynomials of the closed loop transfer function if an appropriate l.s.v.f. control law is employed. If the cancellation of these common factors is made, the "apparent order" of the closed loop system is decreased, thereby allowing us to "match" certain models of lower dynamic order than n, as we will show (by example).

represents a coordinate transformation of the given system to "multi-input companion form" [4], [7]. More precisely, setting $\hat{A} = QAQ^{-1}$, $\hat{B} = QB$, and $\hat{C} = CQ^{-1}$, (1) becomes:

$$\dot{z}(t) = \hat{A}z(t) + \hat{B}u(t), \tag{12a}$$

$$y(t) = \hat{C}z(t) + Du(t) \tag{12b}$$

for $z(t) = Qx(t)$, where $\hat{A} = [\hat{a}_{ij}]$ is a block matrix of the form:

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} & \cdots & \hat{A}_{1m} \\ \hat{A}_{21} & & & \vdots \\ \vdots & & & \vdots \\ \hat{A}_{m1} & \cdots & & \hat{A}_{mm} \end{bmatrix} \tag{13}$$

with $\hat{A}_{ii}$ a $\sigma_i \times \sigma_i$ companion matrix with last (σ_ith) row equal to the row vector $[\hat{a}_{d_i,d_{i-1}+1}, \hat{a}_{d_i,d_{i-1}+2}, \cdots, \hat{a}_{d_i,d_{i-1}}, \hat{a}_{d_i,d_i}]$, and for $i \neq j$, $\hat{A}_{ij}$ is a $\sigma_i \times \sigma_j$ matrix with last (σ_ith) row equal to the row vector $[\hat{a}_{d_i,d_{j-1}+1}, \cdots, \hat{a}_{d_i,d_j}]$ and all other entries identically zero. Furthermore, $\hat{B} = [\hat{b}_{ij}]$ is an $n \times m$ matrix given by:

$$\hat{B} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \hat{b}_{d_1,2} & \hat{b}_{d_1,3} & & \hat{b}_{d_1,m} \\ 0 & 0 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & \hat{b}_{d_2,3} & & \hat{b}_{d_2,m} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \end{bmatrix}. \tag{14}$$

In terms of the coordinate transformation Q, the l.s.v.f. control law (2) becomes: $u(t) = \hat{F}z(t) + Gw(t)$, where $\hat{F} = FQ^{-1}$, and the closed loop transfer matrix, relating $y(s)$ to $w(s)$, namely,

$$\begin{aligned} T_{F,G}(s) &= (C + DF)(sI - A - BF)^{-1}BG + DG \\ &= (\hat{C} + D\hat{F})(sI - \hat{A} - \hat{B}\hat{F})^{-1}\hat{B}G + DG, \end{aligned} \tag{15}$$

and is invariant under Q [6].

We now define $[s^{\sigma}]$ as the $m \times m$ diagonal matrix whose (m) ith diagonal entries equal s^{σ_i}, $\hat{A}_m$ as the $m \times n$ matrix consisting of the (m) ordered d_k rows of $\hat{A}$, $\hat{B}_m$ as the $m \times m$ (nonsingular) matrix consisting of the (m) ordered d_k rows of

$\hat{B}$, and $S(s)$ as the following $n \times m$ matrix of single term monic polynomials, i.e.,

$$S(s) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ s & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ s^{\sigma_1 - 1} & 0 & & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & s^{\sigma_2 - 1} & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & s^{\sigma_m - 1} \end{bmatrix}. \tag{16}$$

In terms of the above definitions, the structure theorem [6] establishes the following relationship for the closed loop transfer matrix:

$$T_{F,G}(s) = (\hat{C} + D\hat{F})S(s)\{[s^{\sigma}] - (\hat{A}_m + \hat{B}_m\hat{F})S(s)\}^{-1}\hat{B}_m G + DG, \tag{17}$$

or, after some rudimentary manipulations,

$$T_{F,G}(s) = \{D\hat{B}_m^{-1}[s^{\sigma}] + (\hat{C} - D\hat{B}_m^{-1}\hat{A}_m)S(s)\}\{G^{-1}\hat{B}_m^{-1}[s^{\sigma}] - G^{-1}(\hat{B}_m^{-1}\hat{A}_m + FQ^{-1})S(s)\}^{-1}. \tag{18}$$

If we now let $R(s) = D\hat{B}_m^{-1}[s^{\sigma}] + (\hat{C} - D\hat{B}_m^{-1}\hat{A}_m)S(s)$, and $P_{F,G}(s) = G^{-1}\hat{B}_m^{-1}\{[s^{\sigma}] - (\hat{A}_m + \hat{B}_m FQ^{-1})S(s)\}$, $T_{F,G}(s)$ can be written more concisely as:

$$T_{F,G}(s) = R(s)P_{F,G}^{-1}(s). \tag{19}$$

It should be noted that (18) represents an extension of (9) to the multivariable case.

It is now immediately apparent, by inspection of the above expression for $T_{F,G}(s)$, that $R(s)$, analogous to $r(s)$ in the scalar case, is unaffected by l.s.v.f. Furthermore, on closer inspection of $P_{F,G}(s)$, an $m \times m$ nonsingular polynomial matrix, we note that the highest degree s-term in each (ith) column of $P_{F,G}(s)$ is s^{σ_i}, a condition which also held for $P(s)$. We thus conclude that the (m) ordered integers σ_i which will be defined as the *controllability indices* of the system (or $P(s)$) are unaltered by l.s.v.f. Furthermore, the coefficients of s^{σ_i} in each (ith) column $P_{F,G}(s)$ equal the elements of each (ith) column of $G^{-1}\hat{B}_m^{-1}$, a nonsingular matrix which can be completely and arbitrarily specified via G. The nonsingularity of the $m \times m$ matrix consisting of the coefficients of the highest degree s-terms in each column of a polynomial matrix (such as $P_{F,G}(s)$) is referred to as the condition that the matrix is "column proper";[2] i.e., $P_{F,G}(s)$ is therefore column proper for all admissible $\{F, G\}$ pairs. Now, we note that the remaining (lower degree) s-terms comprising $P_{F,G}(s)$ can be completely and arbitrarily specified via F. Furthermore,

[2] In general, any $m \times m$ polynomial matrix $P(s)$ is called *column proper* if and only if the $m \times m$ constant matrix consisting of the coefficients of the highest degree s term (or terms) in each column of $P(s)$ is nonsingular. The reader is referred to [8] for a more thorough discussion of this point.

since no additional independent l.s.v.f. quantities can be defined, and the quantities noted are sufficient to characterize the entire class of l.s.v.f. systems of the form (12) for all possible feedback pairs $\{\hat{F}, G\}$, we conclude that:

(i) $R(s)$,

(ii) *the* (m) *ordered controllability indices,* σ_i, *and*

(iii) *the fact that* $P_{F,G}(s)$ *is column proper*

are unaltered by l.s.v.f. and represent a complete and concise *characterization* of the system under l.s.v.f.

It should be noted that unlike the scalar case, neither $R(s)$ nor the ordering of the controllability indices are, in general, unique, since the "multi-input companion form" $\{\hat{A}, \hat{B}\}$ for any given pair $\{A, B\}$ is not unique, but depends on the particular coordinate transformation matrix Q. Once a coordinate system has been fixed, however, both $R(s)$ and the ordering of the σ_i are unique. This nonuniqueness of the characterization does not present any difficulties in application to the question of exact model matching, however, as we shall see in the next section.

5. Exact model matching. We can now employ the characterization developed in the previous section to resolve the question of exact model matching posed in § 2. In order to establish a theorem which provides a constructive answer to this question, we shall require certain fundamental definitions involving polynomial matrices, which we extract directly from MacDuffee [9].

In particular, if three polynomial matrices satisfy the relationship: $P(s) = R(s)H(s)$, then $H(s)$ ($R(s)$) is called a *right* (*left*) *divisor* of $P(s)$, and $P(s)$ is called a *left* (*right*) *multiple* of $H(s)$ ($R(s)$). A *greatest common right divisor* (g.c.r.d.), $G(s)$, of two matrices, $P(s)$ and $R(s)$, is a common right divisor which is a left multiple of every common right divisor of $P(s)$ and $R(s)$. A *greatest common left divisor* (g.c.l.d.) of two polynomial matrices is similarly defined. Finally, a *unimodular matrix* is any polynomial matrix whose determinant is a nonzero scalar (independent of s).

We now state a rather obvious result in view of the above definitions, namely if $T_m(s)$ is any $p \times m$ transfer matrix, then it can be "factored" as

$$T_m(s) = R_m(s)P_m^{-1}(s), \tag{20}$$

where $R_m(s)$ and $P_m(s)$ are $p \times m$ and $m \times m$ (nonsingular) polynomial matrices, respectively, with I_m a g.c.r.d. of $R_m(s)$ and $P_m(s)$. Loosely speaking, the polynomial matrices might be called "relatively prime," although this description should be avoided in the matrix case [8]. The relationship (20) can readily be established by first representing $T_m(s)$ as $N(s)/\Delta(s)$ or as $N(s)[\Delta(s)I_m]^{-1}$, where $N(s)$ is a $p \times m$ polynomial matrix and $\Delta(s)$ is the least common denominator of all (pm) entries of $T_m(s)$, and then "factoring" any nonunimodular g.c.r.d. $G(s)$ from both $N(s)$ and $\Delta(s)I_m$, i.e., once a $G(s)$ has been found, we let $R_m(s) = N(s)G^{-1}(s)$ and $P_m(s) = \Delta_m(s)I_mG^{-1}(s)$, thus establishing the desired result. It should be noted that this procedure represents an extension, to the matrix case, of the procedure of "canceling" common pole-zero factors in a scalar transfer function to reduce the function to the quotient of two relatively prime polynomials. An algorithm for obtaining a g.c.r.d. of two polynomial matrices can be found in [8] or [9].

We now state and establish the main result of this paper.

THEOREM. *Consider the system* (1) *with* $\{A, B\}$ *controllable, and* B *of full rank* $m \leqq n$, *and* $T_m(s) = R_m(s)P_m^{-1}(s)$, *a* $p \times m$ *transfer matrix with* I_m *a g.c.r.d. of* $R_m(s)$ *and* $P_m(s)$. *There exists an l.s.v.f. pair* $\{F, G\}$, *with* G *nonsingular, which satisfies the relationship*:

$$\begin{aligned} T_{F,G}(s) &= (C + DF)(sI - A - BF)^{-1}BG + DG = R(s)P_{F,G}^{-1}(s) \\ &= R_m(s)P_m^{-1}(s) = T_m(s) \end{aligned} \tag{21}$$

if and only if for some nonsingular polynomial matrix $H(s)$ *the following three conditions hold*:

(i) $R_m(s)$ *is a left divisor of* $R(s)$, i.e., $R(s) = R_m(s)H(s)$;

(ii) *the* (m) *ordered* σ_i *of* $P_m(s)H(s)$ *are identical to those of* $P_{F,G}(s)$;

(iii) $P_m(s)H(s)$ *is column proper.*

Remark. Before establishing this theorem, it should be noted that wherever the system is "left invertible", $H(s)$ (if it exists) is unique and can readily be determined by employing condition (i) above as we shall show. If the system is not "left invertible", however, there is, as yet, no general algorithm for finding an appropriate $H(s)$ or even determining if one exists. The question of exact model matching via l.s.v.f. is therefore *completely* resolved here only when the given system is "left invertible".

Proof. The proof of sufficiency is constructive, and simply involves equating $P_m(s)H(s)$ to $P_{F,G}(s)$ once an appropriate $H(s)$ has been found, i.e., by (i) and (21), we have: $R(s)P_{F,G}^{-1}(s) = R_m(s)H(s)[P_m(s)H(s)]^{-1}$. We thus set

$$P_m(s)H(s) = P_{F,G}(s) = G^{-1}\hat{B}_m^{-1}([s^\sigma] - (\hat{A}_m + \hat{B}_m F Q^{-1})S(s)), \tag{22}$$

where $\hat{B}_m$, $\hat{A}_m$, $[s^\sigma]$, Q and $S(s)$ are known. By (iii), the $m \times m$ matrix E, consisting of the coefficients of the highest degree s-terms in each column of $P_m(s)H(s)$, is nonsingular. Therefore, equating E to $G^{-1}\hat{B}_m^{-1}$ produces the appropriate G; i.e., $G^{-1}\hat{B}_m^{-1} = E$, or

$$G = (E\hat{B}_m)^{-1}. \tag{23}$$

A corresponding F can then be determined by first premultiplying (22) by $E^{-1} = \hat{B}_m G$ and then subtracting $([s^\sigma] - \hat{A}_m S(s))$ from both sides, i.e.,

$$\hat{B}_m G P_m(s)H(s) - ([s^\sigma] - \hat{A}_m S(s)) = -\hat{B}_m \hat{F} S(s), \tag{24}$$

from which $\hat{B}_m\hat{F}$ (and therefore $\hat{F}$ and $F = \hat{F}Q$) can be determined by inspection. Note that this step depends on condition (ii) to produce complete subtraction of the elements of $[s^\sigma]$, resulting in the correct structure for $\hat{B}_m\hat{F}S(s)$. This completes our proof of sufficiency.

We now establish necessity. In particular, if (21) is satisfied,

$$R(s) = R_m(s)P_m^{-1}(s)P_{F,G}(s) = R_m(s)P_m^{+}(s)P_{F,G}(s) \div |P_m(s)|, \tag{25}$$

where $P_m^+(s)$ and $|P_m(s)|$ represent the adjoint and determinant of $P_m(s)$, respectively. Since I_m is a g.c.r.d. of $R_m(s)$ and $P_m(s)$, it follows that [9]:

$$M_1(s)R_m(s) + M_2(s)P_m(s) = I_m \tag{26}$$

for a particular pair $\{M_1(s), M_2(s)\}$ of polynomial matrices. Postmultiplying both sides of (26) by $P_m^+(s)P_{F,G}(s)$, we obtain:

$$M_1(s)R_m(s)P_m^+(s)P_{F,G}(s) + M_2(s)P_m(s)P_m^+(s)P_{F,G}(s) = P_m^+(s)P_{F,G}(s). \tag{27}$$

By (25), we see that $|P_m(s)|$ must divide the product $R_m(s)P_m^+(s)P_{F,G}(s)$, since $R(s)$ is a polynomial matrix. Also, since $P_m(s)P_m^+(s) = |P_m(s)|I_m$, it follows from (27) that $|P_m(s)|$ must divide $P_m^+(s)P_{F,G}(s)$, since it divides both left-side members of (27). We thus conclude that $P_m^{-1}(s)P_{F,G}(s)$ is a (nonsingular) polynomial matrix, which we call $H(s)$, i.e.,

$$P_{F,G}(s) = P_m(s)H(s), \tag{28}$$

and, from (25),

$$R(s) = R_m(s)H(s). \tag{29}$$

Equations (28) and (29) directly imply all three conditions of the theorem, thus establishing necessity and completing the proof of the theorem.

It should be noted that successful employment of the theorem to resolve the question of model matching involves (a) the transformation of the dynamical equations of the system to a "companion form," (b) factoring $T_m(s)$ as the product $R_m(s)P_m^{-1}(s)$, where I_m is a g.c.r.d. of $R_m(s)$ and $P_m(s)$, and (c) the determination of an appropriate $H(s)$, if one exists. As noted earlier, there are algorithms available to perform the first two of these steps. Moreover, also as noted earlier, if the system is "left invertible" or, equivalently, if $R(s)$ has rank m, $H(s)$, if it exists, is uniquely specified via (29) and can easily be determined. Furthermore, *whenever $H(s)$ is unique, the feedback pair $\{F, G\}$ is also unique*, a fact which clearly follows from the theorem proof of sufficiency.

We conclude this section by illustrating the constructive algorithm used to establish sufficiency of the theorem. Therefore, consider the following example.

Example. Consider a system of the form (1), where

$$A = \begin{bmatrix} -4 & -1 & -3 \\ 3 & 1 & 1 \\ 5 & 1 & 3 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ -1 & -1 \end{bmatrix}, \quad C = \begin{bmatrix} 4 & -1 & 4 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$D = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Suppose we are now asked to find a feedback pair $\{F, G\}$, if one exists, such that the closed loop transfer matrix of the system, as given by (15), is equal to

$$T_m(s) = \begin{bmatrix} \dfrac{1}{s+3} & 0 \\ \dfrac{-s}{s+3} & 1 \end{bmatrix}.$$

In order to answer this question, we first "reduce" the system to companion form, i.e., using (10) and (11), we find that

$$\sigma_1 = 2, \quad \sigma_2 = 1, \quad Q = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix},$$

and hence that

$$\hat{A} = QAQ^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ -4 & 0 & -1 \\ -2 & 0 & 0 \end{bmatrix}, \quad \hat{B} = QB = \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

and

$$\hat{C} = CQ^{-1} = \begin{bmatrix} 3 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Furthermore, by inspection,

$$\hat{A}_m = \begin{bmatrix} -4 & 0 & -1 \\ -2 & 0 & 0 \end{bmatrix}, \quad \hat{B}_m = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad S(s) = \begin{bmatrix} 1 & 0 \\ s & 0 \\ 0 & 1 \end{bmatrix}.$$

Therefore,

$$R(s) = D\hat{B}_m^{-1}[s^{\sigma}] + (\hat{C} - D\hat{B}_m^{-1}\hat{A}_m)S(s) = \begin{bmatrix} s+3, & -1 \\ 2, & s \end{bmatrix},$$

and is nonsingular, i.e., $|R(s)| = s^2 + 3s + 2$.

We next factor $T_m(s)$ as $R_m(s)P_m^{-1}(s)$, where I_m is a g.c.r.d. of $R_m(s)$ and $P_m(s)$. For this example, it is clear that the choice

$$R_m(s) = \begin{bmatrix} 1 & 0 \\ -s & 1 \end{bmatrix} \quad \text{and} \quad P_m(s) = \begin{bmatrix} s+3 & 0 \\ 0 & 1 \end{bmatrix}$$

is an appropriate one, since $R_m(s)$ is a unimodular matrix. This also implies that $R_m(s)$ is a left divisor of $R(s)$, i.e., using (29),

$$H(s) = R_m^{-1}(s)R(s) \quad \text{or} \quad H(s) = \begin{bmatrix} s+3 & -1 \\ s^2+3s+2, & 0 \end{bmatrix},$$

which implies that condition (i) of the theorem is satisfied. To determine whether or not the remaining two conditions are satisfied, we compute the product:

$$P_m(s)H(s) = \begin{bmatrix} s^2+6s+9, & -s-3 \\ s^2+3s+2, & 0 \end{bmatrix}.$$

Clearly, $P_m(s)H(s)$ is column proper with

$$\sigma_1 = 2, \quad \sigma_2 = 1 \quad \text{and} \quad E = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}.$$

Therefore, for this example, there does exist a (unique) feedback pair $\{F, G\}$, which produces $T_m(s)$. In particular, G is readily obtained via (23), i.e.,

$$G = (E\hat{B}_m)^{-1} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix},$$

and F is obtained via (24), i.e.,

$$\hat{B}_m\hat{F}S(s) = \begin{bmatrix} -3s + 2, & 1 \\ 3s + 9 & -3 \end{bmatrix} \quad \text{or} \quad \hat{F}S(s) = \begin{bmatrix} -6s - 7, & 4 \\ 3s + 9 & -3 \end{bmatrix}.$$

Therefore, by inspection,

$$\hat{F} = \begin{bmatrix} -7 & -6 & 4 \\ 9 & 3 & -3 \end{bmatrix} \quad \text{and} \quad F = \hat{F}Q = \begin{bmatrix} -13 & 4 & -11 \\ 12 & -3 & 12 \end{bmatrix}.$$

This example also serves to illustrate the fact (noted earlier) that the model $T_m(s)$ need not be of the same dynamic order as the given system, i.e., in this example, $T_m(s)$ can clearly be "realized" by a first order state form system, although the given system is third order.

6. Concluding remarks. We have presented a complete and concise characterization of linear, time-invariant, multivariable systems under linear state variable feedback, and employed the characterization to resolve constructively a question of practical importance in control system design, namely the question of exact model matching. The development employed in the general multivariable case was motivated by the rather simple scalar case. The characterization which was developed was shown to depend on a particular "companion form" representation for modeling the dynamical behavior of the system, as well as certain fundamental definitions and relationships involving polynomial matrices. In particular, the notion of a "column proper" polynomial matrix was introduced to complete the particular characterization used.

The assumptions which were made to facilitate the development were rather modest, i.e., system controllability and linear independence of the internal and external inputs were the only significant restrictions placed on the class of systems considered. Thus, models of unspecified dynamical order are included in the treatment.

It should be noted that the procedures employed (l.s.v.f.) for model matching, as formulated in this paper, are the same as those used initially for "decoupling" multivariable systems. With this in mind, it might be expected that, as in the case of "decoupling" research, future studies of model matching might consider the effect of adding dynamics when state feedback is insufficient for exact matching, or when state feedback results in an exact, but unstable, match. Moreover, since the closed loop poles of the system (which matches the model) are equal to the zeros of $|P_m(s)|$ and $|H(s)|$, it is important to choose a "stable" $H(s)$, when such a

choice is possible. However, the question of determining *any* appropriate $H(s)$, when $R(s)$ is not of rank m, is not as yet resolved, and additional studies might well focus on this question also.

REFERENCES

[1] H. ERZBERGER, *Analysis and design of model following control systems by state space techniques*, Proc. 1968 JACC, Ann Arbor, Mich., pp. 572–581.

[2] C. A. WINSOR AND R. J. ROY, *The application of specific optimal control to the design of desensitized model following control systems*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 326–333.

[3] R. E. KALMAN, *The mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[4] W. A. WOLOVICH, *On the stabilization of controllable systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 569–572.

[5] R. W. BROCKETT, *Poles, zeros, and feedback: state space interpretation*, Ibid., AC-10 (1965), pp. 129–135.

[6] W. A. WOLOVICH AND P. L. FALB, *On the structure of multivariable systems*, this Journal, 7 (1969), pp. 437–451.

[7] D. G. LUENBERGER, *Canonical forms for linear multivariable systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 290–293.

[8] W. A. WOLOVICH, *Equivalent representations and realizations of linear multivariable systems*, Engineering Rep. NSF GK-2788/2, Brown University, Providence, R.I., 1970.

[9] C. C. MACDUFFEE, *The Theory of Matrices*, Chelsea, New York, 1956.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

EFFICIENT IMPLEMENTATIONS OF THE POLAK–RIBIÈRE CONJUGATE GRADIENT ALGORITHM*

R. KLESSIG† AND E. POLAK‡

Abstract. Two modifications of the Polak–Ribière conjugate gradient algorithm are presented. Both modifications eliminate the need for an exact minimization at each iteration and both are shown to be convergent. The advantage of the first modification lies in the fact that it takes less time per iteration than the second modification. However, only the second modification is shown to converge n-step quadratically.

1. Introduction. Quite commonly, theoretical algorithms are stated as a recurrence relation of the form $x_{i+1} \in A(x_i)$, $i = 0, 1, 2, \cdots$, where $A(\cdot)$ is a set-valued function and the sequence $\{x_i\}$ converges to a solution point. Also quite commonly, to compute a vector x_{i+1} in the set $A(x_i)$, we must bring in a sub-algorithm which starts out by setting $y_0 = x_i$, and then constructs an infinite sequence $y_0, y_1, y_2, \cdots$ which converges to a point x_{i+1} in $A(x_i)$. Consequently, from a constructive point of view, such an algorithm ($x_{i+1} \in A(x_i)$) is not well-defined because it is doubly infinitely iterative. The problem of implementing an algorithm of the form $x_{i+1} \in A(x_i)$ is that of finding an approximation map $\tilde{A}(\cdot, \cdot)$, possibly depending on a parameter ε, such that, (i) the computation of a point $\tilde{x}_{i+1} \in \tilde{A}(\varepsilon, \tilde{x}_i)$ can be carried out without constructing an infinite sequence $\{y_i\}$, and (ii) when the parameter ε is appropriately manipulated, the sequence $\{\tilde{x}_i\}$ has the same convergence properties as the sequence $\{x_i\}$.

In practice, an implementation of a doubly iterative algorithm is obtained by truncating the construction of the sequence $\{y_i\}$ after a finite number of elements have been obtained. A theoretical basis for this practice in certain cases can be found in [1, §1.3] and in [7]. From the results in [1], as well as from empirical knowledge, it is clear that if the construction of the sequence $\{y_i\}$ is terminated too early, convergence or rate of convergence for the sequence $\{\tilde{x}_i\}$ may be lost, while if the construction of the y_i is allowed to continue for too long, the computation becomes unduly expensive. Thus, the problem of constructing an efficient implementation is far from trivial.

In this paper we shall present two efficient implementations of the Polak–Ribière conjugate gradient algorithm [2] which was introduced in 1969 and independently by B. T. Polyak [8]. This *theoretical* algorithm solves the following problem:

$$\min \{f(z) | z \in \mathbb{R}^n\}, \tag{1.1}$$

where $f: \mathbb{R}^n \to \mathbb{R}^1$ is strictly convex and twice continuously differentiable.

* Received by the editors October 13, 1970, and in final revised form October 15, 1971. This work was sponsored by the National Aeronautics and Space Administration under Grant NGL-05-003-016 (Sup. 7).

† Bell Telephone Laboratories, Holmdel, New Jersey 07716.

‡ Department of Electrical Engineering and Computer Sciences and Electronics Research Laboratory, University of California, Berkeley, California 94720.

We now state this algorithm for future reference.

1.2. THE POLAK–RIBIÈRE CONJUGATE GRADIENT ALGORITHM [2].

Step 0. Select a $z_0 \in \mathbb{R}^n$.

Step 1. If $\nabla f(z_0) = 0$, stop; else set $i = 0$, set $g_0 = h_0 = -\nabla f(z_0)$ and go to Step 2.

Step 2. Compute $\lambda_i \geqq 0$ such that

$$f(z_i + \lambda_i h_i) = \min_{\lambda \geqq 0} f(z_i + \lambda h_i). \tag{1.3}$$

Step 3. Set

$$z_{i+1} = z_i + \lambda_i h_i. \tag{1.4}$$

Step 4. Compute $\nabla f(z_{i+1})$.

Step 5. If $\nabla f(z_{i+1}) = 0$, stop; else, set

$$g_{i+1} = -\nabla f(z_{i+1}), \tag{1.5}$$

$$\gamma_i = \langle g_{i+1} - g_i, g_{i+1} \rangle / \|g_i\|^2, \tag{1.6}$$

$$h_{i+1} = g_{i+1} + \gamma_i h_i, \tag{1.7}$$

set $i = i + 1$ and go to Step 2.

The following convergence result was established in [2].

1.8. THEOREM (Polak–Ribière [2]). *If there exists* $0 < m \leqq M < \infty$ *such that*

$$m\|y\|^2 \leqq \langle y, H(z)y \rangle \leqq M\|y\|^2 \quad \text{for all } y, z \in \mathbb{R}^n, \tag{1.9}$$

where $H(z) \equiv \partial^2 f(z)/\partial z^2$, *then there exists a* $\rho \in (0, 1)$ *such that the sequences* $\{g_i\}$ *and* $\{h_i\}$ *constructed by* (1.2), *in the process of solving* (1.1), *satisfy*

$$\langle g_i, h_i \rangle \geqq \rho \|g_i\| \, \|h_i\|, \qquad i = 0, 1, 2, \cdots, \tag{1.10}$$

and the sequence $\{z_i\}$ *converges to* $\hat{z}$, *the unique minimizer of* $f(\cdot)$.

The operation in Algorithm 1.2 which requires us to use an infinite subprocedure is the minimization on a half-line in (1.3). From this point of view, relation (1.10) describes a very important property of the Polak–Ribière algorithm, for it makes the algorithm rather insensitive to an accumulation of errors in the approximate calculation of the minimum in (1.3), provided that the approximation is carried out intelligently, as we shall see in the next section.

In 1970, A. Cohen [4] established a bound on the rate of convergence for the theoretical algorithm, Algorithm 1.2, modified to reinitialize as shown below.

1.11. DEFINITION. Let v be an integer satisfying $v \geqq n$ and let $I_v = \{0, v, 2v, \cdots\}$. For $i = 0, 1, 2, \cdots$ let $\omega(i) = 0$ if $i \in I_v$ and let $\omega(i) = 1$ otherwise. Suppose that Algorithm 1.2 is modified by replacing (1.6) with

$$\gamma_i = \omega(i + 1)\langle g_{i+1} - g_i, g_{i+1} \rangle / \|g_i\|^2. \tag{1.12}$$

We call the resulting algorithm the *Polak–Ribière algorithm with reinitialization.*

[1] Note that under this assumption the function $f(\cdot)$ has a minimum which is achieved at a unique minimizer $\hat{z}$.

1.13. THEOREM (Cohen [4]). *Suppose that* $\{z_i\}_{i=0}^{\infty}$ *is a sequence constructed by the Polak–Ribière algorithm with reinitialization in solving problem* (1.1). *If* $f(\cdot)$ *is three times continuously differentiable,* (1.9) *holds, and* $z_i \to \hat{z}$ *as* $i \to \infty$, *then there exist an integer* $k \geqq 0$ *and a constant* $q \in (0, \infty)$ *such that*

$$\|z_{i+n} - \hat{z}\| \leqq q\|z_i - \hat{z}\|^2 \quad \text{for all } i \geqq k, \quad i \in I_v. \tag{1.14}$$

In the next two sections we shall construct implementations for Algorithm 1.2, the first of which preserves the relation (1.10), while the second one preserves relations (1.10) and (1.14). Our main theoretical results are given in Theorems 2.11, 2.53, 3.13 and 3.95. Convergence is established in the first two theorems, while n-step quadratic rate of convergence is obtained in the last two.

2. A convergent implementation of the Polak–Ribière algorithm. In this section we shall construct a convergent implementation of the Polak–Ribière algorithm. In the next section we shall modify this implementation so as to ensure that the relation (1.14) is satisfied.

2.1. IMPLEMENTATION OF POLAK–RIBIÈRE ALGORITHM I.

Step 0. Select a $z_0 \in \mathbb{R}^n$ and parameters $\delta_0 \in (0, 1)$, $\rho_0 \in (0, 1)$, $\beta \in (0, 1)$, $\beta' \in (0, 1)$, $\beta'' \in (0, 1)$.[2]

Step 1. Set $g_0 = h_0 = -\nabla f(z_0)$; set $i = 0$.

Step 2. If $g_0 = 0$, stop; else, go to Step 3.

Step 3. Set $z = z_i$, $h = (1/\|h_i\|)h_i$.

Step 4. Define $\theta\colon \mathbb{R}^1 \to \mathbb{R}^1$ by

$$\theta(x) = f(z + xh) - f(z). \tag{2.2}$$

Comment. To compute the step size, we shall apply to $\theta(\cdot)$ several iterations of a secant method which uses a step length formula due to Armijo [5]. The exact number of iterations required will be determined by the test in Step 13.

Step 5. Set $x_0 = 0$, $l = 0$.

Step 6. Compute

$$\theta'(x_l) = \langle \nabla f(z + x_l h), h \rangle. \tag{2.3}$$

Step 7. If $\theta'(x_l) = 0$, set $x = x_l$ and go to Step 15; else set $\eta(x_l) = \theta'(x_l)$ and go to Step 8.

Step 8. Compute the smallest nonnegative integer, $j(x_l)$, which satisfies

$$\theta(x_l - \beta^{j(x_l)}\eta(x_l)) - \theta(x_l) + (\beta^{j(x_l)}/3)\eta(x_l)\theta'(x_l) \leqq 0. \tag{2.4}$$

Step 9. Set $x = x_l - \beta^{j(x_l)}\eta(x_l)$.

Step 10. Compute $\nabla f(z + xh)$.

Step 11. If $\nabla f(z + xh) = 0$, set $z_{i+1} = z + xh$ and stop; else, go to Step 12.

Step 12. Compute $\theta'(x)$ according to (2.3) and set $\cos = \theta'(x)/\|\nabla f(z + xh)\|$.

Step 13. If $|\cos| \leqq \delta_i$, go to Step 15; else, set $x_{l+1} = x$ and go to Step 14.

[2] The authors have found $\delta_0 = \cos 85°$, $\rho_0 = \cos 5°$, $\beta = 0.6$, $\beta' = \beta'' = 0.8$ to be a good choice in a number of problems.

Step 14. Set $l = l + 1$, set $\eta(x_l) = [(x_l - x_{l-1})/(\theta'(x_l) - \theta'(x_{l-1}))]\theta'(x_l)$ and go to Step 8.[3]

Step 15. Set

$$z_{i+1} = z + xh, \tag{2.5}$$

$$g_{i+1} = -\nabla f(z + xh), \tag{2.6}$$

$$\gamma_i = \langle g_{i+1} - g_i, g_{i+1}\rangle / \|g_i\|^2, \tag{2.7}$$

$$h_{i+1} = g_{i+1} + \gamma_i h_i. \tag{2.8}$$

Step 16. If $\langle g_{i+1}, h_{i+1}\rangle \geqq \rho_i \|g_{i+1}\| \, \|h_{i+1}\|$, set $\rho_{i+1} = \rho_i$, $\delta_{i+1} = \delta_i$, and go to Step 17; else, set $\rho_{i+1} = \beta''\rho_i$, $\delta_{i+1} = \beta'\delta_i$, and go to Step 17.

Step 17. Set $i = i + 1$ and go to Step 3.

2.9. LEMMA. *Suppose that* (1.9) *holds and that* $\nabla f(z + xh) \neq 0$ *for all* $x \in \mathbb{R}^1$. *Then Algorithm* 2.1 *cannot cycle indefinitely in the loop contained between Steps* 8 *and* 14 (*i.e., it can jam up at a point* $z = z_i$ *only if the minimizer* $\hat{z}$ *of* $f(\cdot)$ *is on the line* $\{z' | z' = z + xh, x \in \mathbb{R}^1\}$).

Proof. Since (1.9) is satisfied, $\theta(\cdot)$ has a minimum on $\mathbb{R}^1$. Suppose that $\hat{x}$ is the minimizer of $\theta(\cdot)$. Then, since $\nabla f(z + \hat{x}h) \neq 0$, we have

$$\theta'(\hat{x})/\|\nabla f(z + \hat{x}h)\| = 0.$$

Consequently, by continuity, it follows that there exists an $\varepsilon_i > 0$ such that for a given $\delta_i > 0$,

$$|\theta'(x)/\|\nabla f(z + xh)\|| \leqq \delta_i, \tag{2.10}$$

for all $\|x - \hat{x}\| \leqq \varepsilon_i$. But $x_l \to \hat{x}$ as $l \to \infty$ (see Appendix) and hence there exists a finite integer k such that $\|x_l - \hat{x}\| \leqq \varepsilon_i$ for all $l \geqq k$. Therefore, (2.10) is satisfied for $x = x_l$, for all $l \geqq k$.

2.11. THEOREM. *Suppose that* (1.9) *holds and consider the sequences* $\{z_i\}$, $\{g_i\}$, $\{h_i\}$, $\{\rho_i\}$ *constructed by Algorithm* 2.1 *in the process of solving problem* (1.1). *If there exists a* $\rho \in (0, 1)$ *such that* $\rho_i \geqq \rho$ *for* $i = 0, 1, 2, \cdots$, *in the test in Step* 16 *of Algorithm* 2.1, *i.e.,* $\langle g_i, h_i\rangle \geqq \rho \|g_i\| \, \|h_i\|$ *for* $i = 0, 1, 2, \cdots$, *then either the algorithm jams up at a point* $z = z_k$ *and* (*Steps* 5 *to* 14) *constructs a sequence* $\{x_l\}_{l=0}^{\infty}$ *such that* $(z + x_l h) \to \hat{z}$, *as* $l \to \infty$, *where* $\hat{z}$ *is the unique minimizer of* $f(\cdot)$ *over* $\mathbb{R}^n$, *or* $\{z_i\}$ *is infinite and* $z_i \to \hat{z}$ *as* $i \to \infty$.

Proof. In view of Lemma 2.9, the first part of the theorem is trivial. Hence, let us assume that the sequence $\{z_i\}$ is infinite.

---

[3] It is shown in the Appendix that $x_l \to \hat{x}$, the minimizer of $\theta(\cdot)$, superlinearly. Of course, if $\theta(\cdot)$ is not strictly convex, this subprocedure might fail. However, this can be avoided, for example, by replacing $\eta(x_l)$ by 1 whenever $(x_l - x_{l-1})/(\theta'(x_l) - \theta'(x_{l-1}))$ is nonpositive, too large in magnitude, or undefined. Even with this modification, the superlinear convergence is retained if $\hat{x}$ is a strong local minimum. In any event, the results that follow only require that the specific subprocedure used for one-dimensional minimization be such that $x_l \to \hat{x}$, that x_1 be constructed as in the present algorithm, and that $\{\theta(x_l)\}$ decrease monotonically. For example, starting with x_1, one can continue with the Davidon cubic interpolation method. (See [13].)

[4] Similar results can also be found in [9] and [10].

We shall now compute a bound on $j(0) = j(x_0)$ for (2.4). Making use of Taylor's formula for second order expansions, of (1.9), and of the $\rho > 0$ which we have assumed to exist, we obtain (cf. (2.4) with $x_l = 0, \beta^j(x_l)$ replaced by λ, and θ, θ' replaced by the expressions (2.2) and (2.3))

$$f(z - \lambda\langle \nabla f(z), h\rangle h) - f(z) + \frac{\lambda}{3}\langle \nabla f(z), h\rangle^2$$

$$= \langle \nabla f(z), h\rangle^2 \left[-\frac{2\lambda}{3} + \lambda^2 \int_0^1 (1-t)\langle H(z - t\langle \nabla f(z), h\rangle h)h, h\rangle \, dt \right] \tag{2.12}$$

$$\leqq \frac{\lambda}{2}\langle \nabla f(z), h\rangle^2(-\tfrac{4}{3} + \lambda M). \tag{2.13}$$

Since for $l = 0$ ($x_0 = 0$), $j(0)$ is chosen so as to make the left-hand side of (2.4) nonpositive, we see from (2.12) that $j(0) \leqq \hat{\jmath}$, where $\hat{\jmath}$ is the smallest integer such that $-\frac{4}{3} + \beta^j M \leqq 0$.

Consequently, from (2.4), we obtain (since $\theta(x_0) = 0$ and $\theta(x_{l+1}) < \theta(x_l)$ for $l = 0, 1, 2, \cdots$) that, for some $l \geqq 0$,

$$f(z_{i+1}) - f(z_i) = \theta(x_l - \beta^{j(x_l)}\eta(x_l)) \leqq \theta(x_0 - \beta^{j(x_0)}\theta'(x_0))$$

$$\leqq -\frac{\beta^{j(x_0)}}{3\|h_i\|^2}\langle \nabla f(z_i), h_i\rangle^2 \leqq -\frac{\beta^{\hat{\jmath}}\rho^2}{3}\|\nabla f(z_i)\|^2 < 0, \tag{2.14}$$

$$i = 0, 1, 2, \cdots.$$

Now, because of (1.9), the level set $\{z | f(z) \leqq f(z_0)\}$ is compact, and hence the sequence $\{z_i\}$ must have accumulation points. Suppose that $z_i \to z^*$ as $i \to \infty$ for $i \in K \subset \{0, 1, 2, \cdots\}$, and that $\nabla f(z^*) \neq 0$. Since $f(\cdot)$ is continuously differentiable by assumption, there exists an integer k such that $\|\nabla f(z_i)\|^2 \geqq \|\nabla f(z^*)\|^2/2$ for all $i \geqq k$, $i \in K$. Suppose that i and $i + j$ are two consecutive indices in K, with $i \geqq k$; then, because of (2.14),

$$f(z_{i+j}) - f(z_i) = [f(z_{i+j}) - f(z_{i+j-1})] + \cdots$$

$$+ [f(z_{i+1}) - f(z_i)] \leqq -\frac{\beta^{\hat{\jmath}}\rho^2}{6}\|\nabla f(z^*)\|^2, \tag{2.15}$$

which shows that the sequence $\{f(z_i)\}_{i\in K}$ is not Cauchy. But $\{f(z_i)\}_{i\in K}$ must converge to $f(z^*)$ because $f(\cdot)$ is continuous, and hence we have a contradiction. Consequently, $\nabla f(z^*) = 0$. Since there is only one point $\hat{z}$ in $\mathbb{R}^n$ such that $\nabla f(\hat{z}) = 0$, we conclude that $z_i \to \hat{z}$ as $i \to \infty$, which completes our proof.

We shall now show that there exists indeed a $\rho \in (0, 1)$ such that $\rho_i \geqq \rho$ for $i = 0, 1, 2, \cdots$ in the test in Step 16 of Algorithm 2.1, i.e., that for some $\rho \in (0, 1)$, the sequences $\{g_i\}$, $\{h_i\}$ constructed by Algorithm 2.1 satisfy

$$\langle g_i, h_i\rangle \geqq \rho\|g_i\| \, \|h_i\|, \qquad i = 0, 1, 2, \cdots. \tag{2.16}$$

2.17. LEMMA. *Consider the sequences* $\{g_i\}$, $\{h_i\}$ *and* $\{\delta_i\}$ *constructed by Algorithm* 2.1 *in the process of solving the problem* (1.1) (*see* (2.6), (2.7), (2.8), *and*

the instructions in Steps 12 *and* 13 *of Algorithm* 2.1). *Then*

$$\langle \bar{g}_{i+1}, \bar{h}_{i+1} \rangle \geqq \frac{\|g_{i+1}\|}{\|h_{i+1}\|}(1-\delta_i^2) - \delta_i \quad \text{for } i = 0, 1, 2, \cdots, \tag{2.18}$$

where $\bar{g}_{i+1} = (1/\|g_{i+1}\|)g_{i+1}$ *and* $\bar{h}_{i+1} = (1/\|h_{i+1}\|)h_{i+1}$, $i = -1, 0, 1, 2, \cdots$.

Proof. From (2.8),

$$\langle h_{i+1}, h_i \rangle = \langle g_{i+1}, h_i \rangle + \gamma_i \|h_i\|^2, \tag{2.19}$$

and hence,

$$\begin{aligned} |\gamma_i| &= |\langle h_{i+1}, h_i \rangle - \langle g_{i+1}, h_i \rangle| / \|h_i\|^2 \\ &\leqq (\|h_i\| \, \|h_{i+1}\| + |\langle g_{i+1}, h_i \rangle|) / \|h_i\|^2. \end{aligned} \tag{2.20}$$

Now making use of the fact that by construction (see Steps 12 and 13 of Algorithm 2.1),

$$|\langle \bar{g}_{i+1}, \bar{h}_i \rangle| \leqq \delta_i, \qquad i = 0, 1, 2, \cdots, \tag{2.21}$$

we obtain from (2.20) that

$$|\gamma_i| \leqq (\|h_{i+1}\| + \|g_{i+1}\|\delta_i)/\|h_i\|. \tag{2.22}$$

Also from (2.8), and making use of (2.21) and (2.22), we have

$$\begin{aligned} \langle \bar{g}_{i+1}, \bar{h}_{i+1} \rangle &= (\|g_{i+1}\|^2 + \gamma_i \langle g_{i+1}, h_i \rangle) / \|g_{i+1}\| \, \|h_{i+1}\| \\ &\geqq (\|g_{i+1}\|^2 - |\gamma_i| \, |\langle g_{i+1}, h_i \rangle|) / \|g_{i+1}\| \, \|h_{i+1}\| \\ &\geqq \frac{\|g_{i+1}\|}{\|h_{i+1}\|}(1 - \delta_i^2) - \delta_i, \qquad i = 0, 1, 2, \cdots, \end{aligned} \tag{2.23}$$

and hence we are done.

2.24. DEFINITION. Consider the sequences $\{z_i\}$ and $\{h_i\}$ constructed by Algorithm 2.1 in the process of solving problem (1.1). We define the sequences $\{\lambda_i\}$ and $\{H_i\}$ by the following relations (see (2.5)):

$$z_{i+1} = z_i + \lambda_i h_i, \qquad i = 0, 1, 2, \cdots, \tag{2.25}$$

$$H_i = \int_0^1 \frac{\partial^2}{\partial z^2} f(z_i + t\lambda_i h_i) \, dt, \qquad i = 0, 1, 2, \cdots. \tag{2.26}$$

(Note that when (1.9) is satisfied, $m\|y\|^2 \leqq \langle y, H_i y \rangle \leqq M\|y\|^2$ and $\|H_i\| \leqq M$.)

2.27. LEMMA. *Suppose that* (1.9) *holds, and consider the sequences* $\{g_i\}$, $\{h_i\}$, $\{\delta_i\}$ *and* $\{\lambda_i\}$ *constructed by Algorithm* 2.1 *in the process of solving problem* (1.1). *Then*

$$|\lambda_i| \leqq \frac{1}{m\|h_i\|^2}[\|g_i\|^2(1+\delta_{i-1}^2) + \|g_i\| \, \|h_i\|\delta_{i-1} + \|g_{i+1}\| \, \|h_i\|\delta_i] \tag{2.28}$$

for $i = 1, 2, \cdots$.

Proof. To begin with, note that because of (2.21) and (2.22),

$$
\begin{aligned}
|\langle g_i, h_i\rangle| &= |\,\|g_i\|^2 + \gamma_{i-1}\langle g_i, h_{i-1}\rangle| \\
&\leqq \|g_i\|^2 + |\gamma_{i-1}|\,|\langle g_i, h_{i-1}\rangle| \\
&\leqq \|g_i\|^2 + \frac{1}{\|h_{i-1}\|}(\|h_i\| + \|g_i\|\delta_{i-1})\delta_{i-1}\|g_i\|\,\|h_{i-1}\| \\
&= \|g_i\|^2(1 + \delta_{i-1}^2) + \|g_i\|\,\|h_i\|\delta_{i-1}.
\end{aligned} \tag{2.29}
$$

Now, making use of (2.6), (2.26) and of the Taylor formula for first order expansions, we obtain

$$
\begin{aligned}
-g_{i+1} &= -g_i + \lambda_i\left(\int_0^1 \frac{\partial^2}{\partial z^2} f(z_i + t\lambda h_i)\,dt\right) h_i \\
&= -g_i + \lambda_i H_i h_i, \qquad\qquad i = 0; 1, 2, \cdots.
\end{aligned} \tag{2.30}
$$

Consequently, for $i = 0, 1, 2, \cdots$,

$$
-\langle g_{i+1}, h_i\rangle = -\langle g_i, h_i\rangle + \lambda_i\langle h_i, H_i h_i\rangle, \tag{2.31}
$$

and hence, for $i = 0, 1, 2, \cdots$,

$$
\lambda_i = (\langle g_i, h_i\rangle - \langle g_{i+1}, h_i\rangle)/\langle h_i, H_i h_i\rangle. \tag{2.32}
$$

Finally, making use of (2.32), (2.29), (2.21) and (1.9) (which implies that $m\|h_i\|^2 \leqq \langle h_i, H_i h_i\rangle \leqq M\|h_i\|^2$, and hence, that $m \leqq \|H_i\| \leqq M$), we obtain for $i = 1, 2, \cdots$,

$$
\begin{aligned}
|\lambda_i| &\leqq \frac{1}{m\|h_i\|^2}(|\langle g_i, h_i\rangle| + |\langle g_{i+1}, h_i\rangle|) \\
&\leqq \frac{1}{m\|h_i\|^2}[\|g_i\|^2(1 + \delta_{i-1}^2) + \|g_i\|\,\|h_i\|\delta_{i-1} + \|g_{i+1}\|\,\|h_i\|\delta_i],
\end{aligned} \tag{2.33}
$$

which is the desired result.

2.34. LEMMA. *Suppose that* (1.9) *holds, and consider the sequences* $\{g_i\}$, $\{h_i\}$ *and* $\{\delta_i\}$ *constructed by Algorithm* 2.1 *in the process of solving problem* (1.1). *Then*

$$
\frac{\|h_{i+1}\|}{\|g_{i+1}\|} \leqq 1 + \frac{M}{m}(1 + \delta_{i-1}^2) + \frac{M}{m}\left(\delta_{i-1} + \frac{\|g_{i+1}\|}{\|g_i\|}\delta_i\right)\frac{\|h_i\|}{\|g_i\|}, \qquad i = 1, 2, \cdots. \tag{2.35}
$$

Proof. From (2.7), (2.26), (2.28) and (2.30), we obtain

$$
\begin{aligned}
|\gamma_i| &= \left|\frac{\langle g_{i+1} - g_i, g_{i+1}\rangle}{\|g_i\|^2}\right| = \frac{|\lambda_i|\,|\langle H_i h_i, g_{i+1}\rangle|}{\|g_i\|^2} \\
&\leqq \frac{M\|h_i\|\,\|g_{i+1}\|}{\|g_i\|^2}\cdot\frac{1}{m\|h_i\|^2}[\|g_i\|^2(1 + \delta_{i-1}^2) \\
&\qquad + \|g_i\|\,\|h_i\|\delta_{i-1} + \|g_{i+1}\|\,\|h_i\|\delta_i] \\
&= \frac{M}{m}\left[\frac{\|g_{i+1}\|}{\|h_i\|}(1 + \delta_{i-1}^2) + \frac{\|g_{i+1}\|}{\|g_i\|}\delta_{i-1} + \frac{\|g_{i+1}\|^2}{\|g_i\|^2}\delta_i\right].
\end{aligned} \tag{2.36}
$$

Next, from (2.8) and from (2.36) we obtain

$$\begin{aligned}\|h_{i+1}\| &\leqq \|g_{i+1}\| + |\gamma_i|\,\|h_i\| \\ &\leqq \|g_{i+1}\| + \frac{M}{m}\left[\|g_{i+1}\|(1+\delta_{i-1}^2) + \frac{\|g_{i+1}\|\,\|h_i\|}{\|g_i\|}\delta_{i-1} + \frac{\|g_{i+1}\|^2\|h_i\|}{\|g_i\|^2}\delta_i\right].\end{aligned} \tag{2.37}$$

Since (2.35) follows from (2.37) by inspection, we are done.

2.38. LEMMA. *Suppose that* (1.9) *holds and consider the sequences* $\{g_i\}$ *and* $\{h_i\}$ *constructed by Algorithm* 2.1 *in the process of solving problem* (1.1). *Then*

$$\frac{\|g_{i+1}\|}{\|g_i\|} \leqq 1 + 2\frac{M}{m} \quad \text{for } i = 0, 1, 2, \cdots. \tag{2.39}$$

Proof. First, by the Taylor formula, together with (2.6) and (1.9), we obtain, since by construction $f(z_{i+1}) - f(z_i) < 0$, for $i = 0, 1, 2, \cdots$,

$$\begin{aligned}0 > f(z_{i+1}) - f(z_i) &= f(z_i + \lambda_i h_i) - f(z_i) \\ &= -\lambda_i\langle g_i, h_i\rangle + \lambda_i^2\int_0^1 (1-t)\,\langle h_i, \frac{\partial^2}{\partial z^2} f(z_i + t\lambda_i h_i)h_i\rangle\,dt \\ &\geqq -|\lambda_i|\,|\langle g_i, h_i\rangle| + \lambda_i^2\int_0^1 (1-t)m\|h_i\|^2\,dt \\ &\geqq -|\lambda_i|\,\|g_i\|\,\|h_i\| + \frac{\lambda_i^2}{2}m\|h_i\|^2, \qquad i = 0, 1, 2, \cdots.\end{aligned} \tag{2.40}$$

Consequently, for $i = 0, 1, 2, \cdots$,

$$|\lambda_i| \leqq \frac{2}{m}\frac{\|g_i\|}{\|h_i\|}. \tag{2.41}$$

Next, from (2.30) and making use of the fact that $\|H_i\| \leqq M$,

$$\|g_{i+1}\| \leqq \|g_i\| + |\lambda_i|\,\|H_i\|\,\|h_i\| \leqq \left(1 + \frac{2M}{m}\right)\|g_i\|, \tag{2.42}$$

from which (2.39) follows directly.

2.43. LEMMA. *Suppose that* (1.9) *is satisfied and consider the sequences* $\{g_i\}$, $\{h_i\}$ *and* $\{\delta_i\}$ *constructed by Algorithm* 2.1 *in the process of solving problem* (1.1). *Let* $\mu \in (0, 1)$. *If there exists an integer* N *such that*

$$\delta_i \leqq \delta \triangleq \frac{1}{2}\left(\frac{m}{m+M}\right)\frac{m}{M}\mu \quad \text{for } i = N-1, N, N+1, \cdots, \tag{2.44}$$

then there exists an $L \in (0, \infty)$ *such that*

$$\frac{\|h_i\|}{\|g_i\|} \leqq L \quad \text{for } i = 0, 1, 2, \cdots. \tag{2.45}$$

Proof. First, making use of (2.39) and (2.44) we obtain that

$$\frac{M}{m}\left(\delta_{i-1} + \frac{\|g_{i+1}\|}{\|g_i\|}\delta_i\right) \leqq \frac{M}{m}\left(\delta_{i-1} + \delta_i + \frac{2M}{m}\delta_i\right) \leqq \frac{2M}{m}\left(1 + \frac{M}{m}\right)\delta = \mu < 1 \tag{2.46}$$

for $i = N, N+1, \cdots$.

Next, substituting from (2.46) into (2.35), we obtain, since $\delta_i \in (0, 1)$ for $i = 0, 1, 2, \cdots$,

$$\frac{\|h_{i+1}\|}{\|g_{i+1}\|} \leqq \left[1 + \frac{M}{m}(1 + \delta^2)\right] + \mu \frac{\|h_i\|}{\|g_i\|} \quad \text{for } i = N, N+1, \cdots. \tag{2.47}$$

Let

$$v = 1 + \frac{M}{m}(1 + \delta^2). \tag{2.48}$$

Then, from (2.47), for any $i \in \{N+1, N+2, \cdots\}$, since $\mu \in (0, 1)$, we have

$$\begin{aligned} \frac{\|h_i\|}{\|g_i\|} &\leqq v + \mu \frac{\|h_{i-1}\|}{\|g_{i-1}\|} \\ &\leqq v \sum_{j=0}^{i-N-1} \mu^j + \mu^{i-N} \frac{\|h_N\|}{\|g_N\|} \\ &\leqq v \sum_{j=0}^{\infty} \mu^j + \mu^{i-N} \frac{\|h_N\|}{\|g_N\|} \leqq \frac{v}{1-\mu} + \frac{\|h_N\|}{\|g_N\|}. \end{aligned} \tag{2.49}$$

Now since $\delta_i \in (0, 1]$ for $i = 0, 1, 2, \cdots$, and because of (2.35) and (2.39),

$$\frac{\|h_{i+1}\|}{\|g_{i+1}\|} \leqq \left(1 + \frac{2M}{m}\right) + \frac{2M}{m}\left(1 + \frac{M}{m}\right)\frac{\|h_i\|}{\|g_i\|} \quad \text{for } i = 0, 1, 2, \cdots. \tag{2.50}$$

Consequently, since $h_0 = g_0$ (and since $1 + 2M/m > 1$),

$$\begin{aligned} \frac{\|h_i\|}{\|g_i\|} &\leqq \left(1 + \frac{2M}{m}\right) \sum_{j=0}^{i-1} \left[\frac{2M}{m}\left(1 + \frac{M}{m}\right)\right]^j + \left[\frac{2M}{m}\left(1 + \frac{M}{m}\right)\right]^i \\ &\leqq \left(1 + \frac{2M}{m}\right) \sum_{j=0}^{i} \left[\frac{2M}{m}\left(1 + \frac{M}{m}\right)\right]^j \leqq \left(1 + \frac{2M}{m}\right) \sum_{j=0}^{N} \left[\frac{2M}{m}\left(1 + \frac{M}{m}\right)\right]^j \\ &\qquad \text{for } i = 0, 1, 2, \cdots, N. \end{aligned} \tag{2.51}$$

Consequently, combining (2.51) and (2.49), we obtain, for $i = 0, 1, 2, \cdots$,

$$\frac{\|h_i\|}{\|g_i\|} \leqq \frac{v}{1-\mu} + \left(1 + \frac{2M}{m}\right) \sum_{j=0}^{N} \left[\frac{2M}{m}\left(1 + \frac{M}{m}\right)\right]^j \triangleq L < \infty. \tag{2.52}$$

2.53. THEOREM. *Suppose that* (1.9) *holds and consider the sequence* $\{\rho_i\}$ *constructed by Algorithm* 2.1 *in the process of solving problem* (1.1). *Then there exists a* $\rho \in (0, 1]$ *such that* $\rho_i \geqq \rho$ *for* $i = 0, 1, 2, \cdots$.

Proof. If Algorithm 2.1 jams up after a finite number of iterations, then the ρ_i are obviously bounded. Hence we only need to consider the case when the sequence $\{\rho_i\}$ is infinite.

Let $\bar{g}_i = (1/\|g_i\|)g_i$, $\bar{h}_i = (1/\|h_i\|)h_i$, $i = 0, 1, 2, \cdots$, and suppose that $\delta_i \to 0$ as $i \to \infty$. We shall show that this leads to a contradiction.

Since $\delta_i \to 0$ as $i \to \infty$, the conditions of Lemma 2.43 are satisfied, and hence there exists an $L \in (0, \infty)$ such that $\|h_i\|/\|g_i\| \leqq L$ for $i = 0, 1, 2, \cdots$. Therefore,

from (2.18), we obtain

$$\begin{aligned}\langle \bar{g}_i, \bar{h}_i \rangle &\geqq \frac{\|g_i\|}{\|h_i\|}(1 - \delta_{i-1}^2) - \delta_{i-1} \\ &\geqq \frac{1}{L}(1 - \delta_{i-1}^2) - \delta_{i-1}, \qquad i = 1, 2, \cdots.\end{aligned} \tag{2.54}$$

Consequently, since $\delta_i \to 0$ as $i \to \infty$, there exists an integer $N' \geqq 0$ such that

$$\langle \bar{g}_i, \bar{h}_i \rangle \geqq \frac{1}{2L} \triangleq \hat{\rho} > 0 \quad \text{for } i = N', N' + 1, \cdots. \tag{2.55}$$

Now Step 16 of Algorithm 2.1 implies that $\rho_{N'} \geqq (\beta'')^{N'}\rho_0$, and therefore (2.55) and Step 16 of Algorithm 2.1 imply that for $i \geqq N'$, $\rho_i \geqq (\beta'')^j\rho_0 > 0$, where j is the smallest positive integer not less than N' such that $(\beta'')^j\rho_0 \leqq \hat{\rho}$. But $\delta_i \to 0$ as $i \to \infty$ if and only if $\rho_i \to 0$ as $i \to \infty$, according to the instruction in Step 16 of Algorithm 2.1. Hence we have a contradiction, and therefore $\{\delta_i\}$ does not converge to zero. Consequently, $\{\rho_i\}$ does not converge to zero; therefore the existence of a $\rho > 0$ such that $\rho_i \geqq \rho$ for $i = 0, 1, 2 \cdots$ has been established.

Consequently, the assumptions of Theorem 2.11 are satisfied by Algorithm 2.1.

3. An *n*-step quadratically converging implementation of the Polak–Ribière algorithm. The convergent implementation, Algorithm 2.1, has the very nice feature that it maintains within fixed limits the precision with which the minimization of $\theta(x)$ (see 2.2)) is carried out. This fixed precision is defined by the tests in Steps 13 and 16, and results from the fact, established in Theorem 2.53, that $\rho_i \geqq \rho > 0$ and $\delta_i \geqq \delta > 0$ for $i = 0, 1, 2, \cdots$. However, if we wish to ensure that the result (1.14) be valid for the sequence $\{z_i\}$ constructed, then we must reinitialize as in Definition 1.11 and make the minimization of $\theta(x)$ (see (2.2)) progressively more precise, as shown in Steps 13, 15, 16 of the algorithm below (cf. Algorithm 2.1).

3.1. Implementation of Polak–Ribière Algorithms II.

Step 0. Select a $z_0 \in \mathbb{R}^n$ and parameters

$$\tilde{\delta}_0 \in (0, 1], \quad \rho_0 \in (0, 1], \quad \beta \in (0, 1), \quad \beta' \in (0, 1), \quad \beta'' \in (0, 1).$$

Step 1. Set $g_0 = h_0 = -\nabla f(z_0)$; set $i = 0$.

Step 2. If $g_0 = 0$, stop; else go to Step 3.

Step 3. Set $z = z_i$, $h = (1/\|h_i\|)h_i$.

Step 4. Define $\theta\colon \mathbb{R}^1 \to \mathbb{R}^1$ by

$$\theta(x) = f(z + xh) - f(z). \tag{3.2}$$

Step 5. Set $x_0 = 0$, set $l = 0$.

Step 6. Compute

$$\theta'(x_l) = \langle \nabla f(z + x_l h), h \rangle. \tag{3.3}$$

Step 7. If $\theta'(x_l) = 0$, go to Step 15; else set $\eta(x_l) = \theta'(x_l)$ and go to Step 8.

Step 8. Compute the smallest nonnegative integer $j(x_l)$ which satisfies

$$\theta(x_l - \beta^{j(x_l)}\eta(x_l)) - \theta(x_l) + \frac{\beta^{j(x_l)}}{3}\theta'(x_l)\eta(x_l) \leqq 0. \tag{3.4}$$

Step 9. Set $x = x_l - \beta^{j(x_l)}\theta'(x_l)$.

Step 10. Compute $\nabla f(z + xh)$.

Step 11. If $\nabla f(z + xh) = 0$, set $z_{i+1} = z + xh$ and stop; else go to Step 12.

Step 12. Compute $\theta'(x)$ according to (3.3) and set $\cos = \theta'(x)/\|\nabla f(z + xh)\|$.

Step 13. If $|\cos| \leqq \delta_i \triangleq \min\{\tilde{\delta}_i, \|g_i\|\}$, go to Step 15; else, set $x_{l+1} = x$ and go to Step 14.

Step 14. Set $l = l + 1$, set

$$\eta(x_l) = \left[\frac{x_l - x_{l-1}}{\theta'(x_l) - \theta'(x_{l-1})}\right]\theta'(x_l)$$

and go to Step 8.

Step 15. Set

$$z_{i+1} = z + xh, \tag{3.5}$$

$$g_{i+1} = -\nabla f(z + xh), \tag{3.6}$$

$$\gamma_i = \omega(i + 1)\langle g_{i+1} - g_i, g_{i+1}\rangle/\|g_i\|^2, \tag{3.7}$$

$$h_{i+1} = g_{i+1} + \gamma_i h_i, \tag{3.8}$$

where $\omega(i + 1)$ is as in Definition 1.11.

Step 16. If $\langle g_{i+1}, h_{i+1}\rangle \geqq \rho_i\|g_{i+1}\|\,\|h_{i+1}\|$, set $\rho_{i+1} = \rho_i$, $\tilde{\delta}_{i+1} = \tilde{\delta}_i$ and go to Step 17; else, set $\rho_{i+1} = \beta''\rho_i$, $\tilde{\delta}_{i+1} = \beta'\tilde{\delta}_i$ and go to Step 17.

Step 17. Set $i = i + 1$ and go to Step 3.

We begin by noting that if $\{z_i\}$ is an infinite sequence constructed by Algorithm 3.1 in the process of solving problem (1.1), then $z_i \to \hat{z}$ as $i \to \infty$, where $\hat{z}$ is the unique minimizer of $f(\cdot)$. To see this, note that for $i \in I_v$ (see Definition 1.11), $h_i = -\nabla f(z_i)$, and from (2.14) with $\rho = 1$,

$$f(z_{i+1}) - f(z_i) \leqq -\frac{\beta^{\hat{j}}}{3}\|\nabla f(z_i)\|^2 \quad \text{for } i \in I_v.$$

Since $f(\cdot)$ is bounded from below and $\{f(z_i)\}$ is a monotically decreasing sequence, we conclude that $\nabla f(z_i) \to 0$ and that $f(z_i) \to f(\hat{z})$ as $i \to \infty$. Since the level sets of $f(\cdot)$ are compact and since $\hat{z}$ is the unique minimizer of $f(\cdot)$, we conclude that $\hat{z}$ must be the unique accumulation point of $\{z_i\}$, i.e., that $z_i \to \hat{z}$ as $i \to \infty$. When the sequence is not infinite, Lemma 2.9 applies.

To establish that (1.14) holds, we proceed essentially as in [4], following a pattern of proof first used by J. Daniel [6] in conjunction with yet another theoretical, conjugate gradient algorithm. Basically, the approach consists in establishing the rate of convergence of our algorithm by means of a suitable comparison with the Newton–Raphson method which uses the recursion formula

$$z_{i+1} = z_i - H(z_i)^{-1}\nabla f(z_i), \qquad i = 0, 1, \cdots, \tag{3.9}$$

in minimizing the twice continuously differentiable function $f(\cdot)$. In (3.9), $H(z_i) = \partial^2 f(z_i)/\partial z^2$, as before. For the purpose of this comparison, we introduce the following sequence of approximating functions.

3.10. Definition. Consider the sequence $\{z_i\}$ generated by Algorithm 3.1 in the process of solving problem (1.1). Then, for $i \in I_v$ (see Definition 1.11) we

define the functions $f_i\colon \mathbb{R}^n \to \mathbb{R}^1$ by

$$f_i(z) = f(z_i) + \langle \nabla f(z_i), z - z_i\rangle + \tfrac{1}{2}\langle z - z_i, H(z_i)(z - z_i)\rangle. \tag{3.11}$$

Since the functions $f_i(\cdot)$ are quadratic, the Polak–Ribière algorithm, Algorithm 1.2, finds their minimizers in at most n iterations.

To ensure that we do not confuse the various sequences constructed in minimizing the $f_i(\cdot)$ with the sequences constructed in minimizing $f(\cdot)$, we shall designate sequences associated with $f_i(\cdot)$, by a subscript i and an overscript j, e.g., $\overset{j}{z}_i$, $\overset{j}{\lambda}_i$, etc. The overscript will be the running index.

3.12. DEFINITION. Consider the sequence $\{z_i\}$ generated by Algorithm 3.1 in the process of solving problem (1.1). For $i \in I_v$, we shall denote by $\overset{j}{z}_i, \overset{j}{g}_i, \overset{j}{h}_i, \overset{j}{\lambda}_i, \overset{j}{\gamma}_i$, $j = 0, 1, 2, \cdots, n$, the quantities constructed by the Polak–Ribière algorithm, Algorithm 1.2, in the process of minimizing the function $f_i(\cdot)$, with Algorithm 1.2 being initialized with $\overset{0}{z}_i = z_i$. Note that for $i \in I_v$, $\overset{0}{h}_i = \overset{0}{g}_i = h_i = g_i = -\nabla f(\overset{0}{z}_i)$.

3.13. THEOREM. *Suppose that the function $f(\cdot)$ in* (1.1) *is three times continuously differentiable and that* (1.9) *holds. Consider the sequences $\{z_i\}$, $\{h_i\}$ and $\{\lambda_i\}$ constructed by Algorithm* 3.1 *in the process of solving problem* (1.1) *and let $\hat{z}$ be the limit point of $\{z_i\}$. If there exists a $q \in (0, \infty)$ and an integer N' such that*

$$\|\lambda_{i+j}h_{i+j} - \overset{j}{\lambda}_i \overset{j}{h}_i\| \leqq q\|z_i - \hat{z}\|^2 \quad \text{for all } i \geqq N',\ i \in I_v,\ j = 0, 1, 2, \cdots, v(i) - 1, \tag{3.14}$$

where $v(i) \leqq n$ is such that $\overset{v(i)}{h}_i = 0$, $\overset{v(i)-1}{h}_i \neq 0$, then there exists a $\hat{q} \in (0, \infty)$ and an integer N'' such that

$$\|z_{i+n} - \hat{z}\| \leqq \hat{q}\|z_i - \hat{z}\|^2 \quad \text{for all } i \geqq N'',\ i \in I_v.^5 \tag{3.15}$$

Proof. Since the functions $f_i(\cdot)$, $i \in I_v$, are quadratic, the point $\overset{v(i)}{z}_i$, $i \in I_v$, (see Definition 3.12) minimizes the function $f_i(\cdot)$ over $\mathbb{R}^n$. Since $H(z_i)$ in (3.11) is nonsingular because of (1.9), the Newton–Raphson method (3.9) can be applied to the minimization of $f_i(\cdot)$, and, since $f_i(\cdot)$ is quadratic, it computes $\overset{v(i)}{z}_i$ (the unique minimizer of $f_i(\cdot)$) in one iteration. Thus,

$$\begin{aligned} \overset{v(i)}{z}_i &= \overset{0}{z}_i - H(z_i)^{-1}\nabla f_i(\overset{0}{z}_i) \\ &= z_i - H(z_i)^{-1}\nabla f(z_i), \qquad i \in I_v \end{aligned} \tag{3.16}$$

(compare the second part of (3.16) with (3.9)!).

Let $a(z) \triangleq z - H(z)^{-1}\nabla f(z)$ denote the Newton–Raphson iteration function. Then, since by assumption $f(\cdot)$ is three times continuously differentiable, there exists an $\varepsilon > 0$ and a $q' \in (0, \infty)$ such that

$$\|a(z) - \hat{z}\| \leqq q'\|z - \hat{z}\|^2 \tag{3.17}$$

for all z such that $\|z - \hat{z}\| < \varepsilon$, where $\hat{z}$ is the minimizer of $f(\cdot)$ (see [1, Theorem (6.2.17)]). Consequently, since $z_i \to \hat{z}$ and because of the second part of (3.16),

[5] We shall show in Theorem 3.95 that the sequences $\{z_i\}$, $\{h_i\}$ and $\{\lambda_i\}$, constructed by Algorithm 3.1, do indeed satisfy the assumptions of Theorem 3.13.

there exists an integer $\tilde{N}'' > N'$ such that

$$\|\overset{v(i)}{z}_i - \hat{z}\| \leqq q'\|z_i - \hat{z}\|^2 \quad \text{for all } i \geqq N'', \quad i \in I_v. \tag{3.18}$$

Now, since $\overset{0}{z}_i = z_i$, $i \in I_v$, because of (3.14), for $i \in I_v$,

$$\begin{aligned}
\|z_{i+\nu(i)} - \overset{\nu(i)}{z}_i\| &= \|[(z_{i+\nu(i)} - z_{i+\nu(i)-1}) + (z_{i+\nu(i)-1} - z_{i+\nu(i)-2}) + \cdots \\
&\quad + (z_{i+1} - z_i)] - [(\overset{\nu(i)}{z}_i - \overset{\nu(i)-1}{z}_i) + \cdots + (\overset{1}{z}_i - \overset{0}{z}_i)]\| \\
&\leqq \sum_{j=0}^{\nu(i)-1} \|(z_{i+j+1} - z_{i+j}) - (\overset{j+1}{z}_i - \overset{j}{z}_i)\| \qquad (3.19) \\
&= \sum_{j=0}^{\nu(i)-1} \|\lambda_{i+j}h_{i+j} - \overset{j}{\lambda}_i\overset{j}{h}_i\| \\
&\leqq nq\|z_i - \hat{z}\|^2 \quad \text{for all } i \geqq N', \quad i \in I_v.
\end{aligned}$$

Consequently, because of (3.18) and (3.19), and because $z_i \to \hat{z}$ at least linearly (see [1, §6.1]), there exists a $c \in (0, \infty)$ and an integer $N'' \geqq \tilde{N}''$ such that

$$\begin{aligned}
\|z_{i+n} - \hat{z}\| &\leqq c\|z_{i+\nu(i)} - \hat{z}\| \leqq c(\|z_{i+\nu(i)} - \overset{\nu(i)}{z}_i\| + \|\overset{\nu(i)}{z}_i - \hat{z}\|) \\
&\leqq c(nq + q')\|z_i - \hat{z}\|^2 \qquad (3.20) \\
&\triangleq \hat{q}\|z_i - \hat{z}\|^2 \quad \text{for all } i \geqq N'', \quad i \in I_v,
\end{aligned}$$

which completes our proof.

The verification that (3.14) is satisfied by Algorithm 3.1 is quite laborious and requires a number of preliminary results which we shall now establish.

To simplify the statements of the lemmas and theorems to follow, we shall assume from now on without loss of generality that $\nu(i) = n$ for $i = 0, 1, 2, \cdots$, that (1.9) is satisfied, that $f(\cdot)$ is three times continuously differentiable (whether required by a specific lemma or not), and that we are dealing with a set of infinite sequences $\{z_i\}$, $\{g_i\}$, $\{h_i\}$, $\{\lambda_i\}$, $\{\gamma_i\}$, $\{\delta_i\}$, $\{\tilde{\delta}_i\}$ and $\{\rho_i\}$ constructed by Algorithm 3.1 in the process of solving problem (1.1) from a given initial point z_0. Corresponding sequences $\{\overset{j}{z}_i\}$, $\{\overset{j}{g}_i\}$, $\{\overset{j}{h}_i\}$, $\{\overset{j}{\lambda}_i\}$ and $\{\overset{j}{\gamma}_i\}$ are as in Definition 3.12, for $i \in I_v$ and $j = 0, 1, 2, \cdots, n$.

3.21. LEMMA. *There exists a $q_1 \in (0, \infty)$ such that*

$$|\gamma_i| \leqq q_1 \quad \textit{for } i = 0, 1, 2, \cdots. \tag{3.22}$$

Proof. By (1.9), (3.7), (2.30), (2.41) and (2.39),

$$\begin{aligned}
|\gamma_i| &= \omega(i+1)|\langle g_{i+1} - g_i, g_{i+1}\rangle|/\|g_i\|^2 \\
&\leqq |\lambda_i|\,|\langle H_i h_i, g_{i+1}\rangle|/\|g_i\|^2 \qquad (3.23) \\
&\leqq \frac{2}{m}\frac{\|g_i\|}{\|h_i\|}\frac{M\|h_i\|\,\|g_{i+1}\|}{\|g_i\|^2} = \frac{2M}{m}\frac{\|g_{i+1}\|}{\|g_i\|} \\
&\leqq \frac{2M}{m}\left(1 + \frac{2M}{m}\right) \triangleq q_1 \quad \text{for } i = 0, 1, 2, \cdots.
\end{aligned}$$

3.24. LEMMA. *There exists an integer* $N \geqq 0$ *and* q_2, q_3 *in* $(0, \infty)$ *such that*

$$q_3\|g_i\| \leqq \|h_i\| \leqq q_2\|g_i\| \quad \textit{for all } i \geqq N. \tag{3.25}$$

Proof. Making use of (3.23) and (3.8), we conclude that

$$\|h_{i+1}\| \leqq \|g_{i+1}\| + |\gamma_i|\|h_i\| \leqq \|g_{i+1}\|\left(1 + \frac{2M}{m}\frac{\|h_i\|}{\|g_i\|}\right), \qquad i = 0, 1, \cdots.$$

Consequently,

$$\frac{\|h_{i+1}\|}{\|g_{i+1}\|} \leqq 1 + \frac{2M}{m}\left(\frac{\|h_i\|}{\|g_i\|}\right), \qquad i = 0, 1, 2, \cdots.$$

But $h_i = g_i$ for $i \in I_v$, and hence, for $i \in I_v$ and $j = 0, 1, 2, \cdots, v - 1$,

$$\frac{\|h_{i+j}\|}{\|g_{i+j}\|} \leqq \sum_{k=0}^{j}\left(\frac{2M}{m}\right)^k \leqq \sum_{k=0}^{v-1}\left(\frac{2M}{m}\right)^k \triangleq q_2;$$

i.e., the right-hand side of (3.25) holds.

We now establish the second half of (3.25). From (3.8), (3.23) and because by Step 13 of (3.1),

$$|\langle \bar{g}_{i+1}, \bar{h}_i\rangle| \leqq \|g_i\| \quad \text{for } i = 0, 1, 2, \cdots$$

($\bar{g}_i = (1/\|g_i\|)g_i$, $\bar{h}_i = (1/\|h_i\|)h_i$), we obtain

$$\begin{aligned}
\|h_i\|^2 &= \|g_i + \gamma_{i-1}h_{i-1}\|^2 \\
&= \|g_i\|^2 + \gamma_{i-1}^2\|h_{i-1}\|^2 + 2\gamma_{i-1}\langle g_i, h_{i-1}\rangle \\
&\quad + \frac{\langle g_i, h_{i-1}\rangle^2}{\|h_{i-1}\|^2} - \frac{\langle g_i, h_{i-1}\rangle^2}{\|h_{i-1}\|^2} \\
&= \|g_i\|^2 - \frac{\langle g_i, h_{i-1}\rangle^2}{\|h_{i-1}\|^2} + \left(\frac{\langle g_i, h_{i-1}\rangle}{\|h_{i-1}\|} + \gamma_{i-1}\|h_{i-1}\|\right)^2 \\
&\geqq \|g_i\|^2 - \frac{\langle g_i, h_{i-1}\rangle^2}{\|h_{i-1}\|^2} = \|g_i\|^2(1 - \cos^2),
\end{aligned} \tag{3.26}$$

where cos is defined as in Step 12 of Algorithm 3.1. Since $\cos^2 \to 0$ as $i \to \infty$, there exists an integer N such that

$$\|h_i\|^2 \geqq \tfrac{1}{2}\|g_i\|^2 \quad \text{for all } i \geqq N. \tag{3.27}$$

3.28. COROLLARY. *There exists an integer* N *and* q_4, q_5 *in* $(0, \infty)$ *such that*

$$|\lambda_i| \leqq q_4 \quad \textit{for all } i \geqq N \tag{3.29}$$

and

$$\|g_{i+1}\| \leqq q_5\|h_i\| \quad \textit{for all } i \geqq N. \tag{3.30}$$

Proof. The inequality (3.29) follows from (2.41) and (3.25). Next, with N as in (3.25), it follows from (2.30), (1.9), (3.25) and (3.29) that

$$\begin{aligned}
\|g_{i+1}\| &\leqq \|g_{i+1} - g_i\| + \|g_i\| \\
&= |\lambda_i|\|H_ih_i\| + \|g_i\| \\
&\leqq (q_4M + 1/q_3)\|h_i\| \triangleq q_5\|h_i\|, \quad i = N, N + 1, \cdots.
\end{aligned} \tag{3.31}$$

The following two lemmas can be obtained by making use of (1.9), (3.22), (3.25), (3.29), (3.30) and of the fact that $\langle \overset{j+1}{g_i}, \overset{j}{g_i} \rangle = 0$ and that $\langle \overset{j+1}{h_i}, H(z_i)\overset{j}{h_i} \rangle = 0$ for $j = 0, 1, 2, \cdots, n-1$ and $i \in I_v$ (see (6.3.20)–(6.3.31) in [1], or [11]).

3.32. LEMMA. *For $i \in I_v$ and $j = 0, 1, \cdots, n-1$,*

$$\overset{j}{\lambda}_i = \langle \overset{j}{g_i}, \overset{j}{h_i} \rangle / \langle \overset{j}{h_i}, H(z_i)\overset{j}{h_i} \rangle = \|\overset{j}{g_i}\|^2 / \langle \overset{j}{h_i}, H(z_i)\overset{j}{h_i} \rangle, \tag{3.33}$$

$$\begin{aligned} \overset{j}{\gamma}_i &= \langle \overset{j+1}{g_i} - \overset{j}{g_i}, \overset{j+1}{g_i} \rangle / \|\overset{j}{g_i}\|^2 \\ &= -\langle \overset{j+1}{g_i}, H(z_i)\overset{j}{h_i} \rangle / \langle \overset{j}{h_i}, H(z_i)\overset{j}{h_i} \rangle. \end{aligned} \tag{3.34}$$

3.35. LEMMA. *For $i \in I_v$ and $j = 0, 1, 2, \cdots, n-1$, there exist $\bar{q}_1, \bar{q}_2, \bar{q}_3, \bar{q}_4, \bar{q}_5$ in $(0, \infty)$ such that*

$$\bar{q}_3 \|\overset{j}{g_i}\| \leqq \|\overset{j}{h_i}\| \leqq \bar{q}_2 \|\overset{j}{g_i}\|, \tag{3.36}$$

$$\|\overset{j+1}{g_i}\| \leqq \bar{q}_5 \|\overset{j}{h_i}\|, \tag{3.37}$$

$$\overset{j}{\lambda}_i \leqq \bar{q}_4, \tag{3.38}$$

$$|\overset{j}{\gamma}_i| \leqq \bar{q}_1. \tag{3.39}$$

3.40. LEMMA. *There exists an integer N and a $q_6 \in (0, \infty)$ such that*

$$\|g_{i+j}\| \leqq q_6 \|h_i\|, \qquad i \geqq N, \quad j = 0, 1, \cdots, n, \tag{3.41}$$

$$\|h_{i+j}\| \leqq q_6 \|h_i\|, \qquad i \geqq N, \quad j = 0, 1, \cdots, n, \tag{3.42}$$

$$\|\overset{j}{g_i}\| \leqq q_6 \|h_i\|, \qquad i \geqq N, \quad i \in I_v, \quad j = 0, 1, \cdots, n, \tag{3.43}$$

$$\|\overset{j}{h_i}\| \leqq q_6 \|h_i\|, \qquad i \geqq N, \quad i \in I_v, \quad j = 0, 1, \cdots, n. \tag{3.44}$$

Proof. Making use of (3.25) and of (3.30) we obtain, for $i \geqq N$ and $j = 1, 2, \cdots, n$,

$$\|g_{i+j}\| \leqq q_5 \|h_{i+j-1}\| \leqq q_5 q_2 \|g_{i+j-1}\| \leqq q_5 (q_2 q_5)^{j-1} \|h_i\|. \tag{3.45}$$

Combining (3.45) with (3.25), we obtain an inequality of the form of (3.41). Since $\overset{0}{h}_i = h_i = g_i$ for $i \in I_v$ an inequality of the form of (3.43) follows similarly from (3.36) and (3.37); an inequality of the form of (3.42) now follows from (3.25) and (3.41), and an inequality of the form of (3.44) from (3.36) and (3.43). Setting q_6 to be the largest of the constants in these inequalities, we see that the lemma holds.

Notation. We shall denote by $O_l(\cdot)$, $l = 1, 2, 3, \cdots$, functions from $\mathbb{R}^1$ into $\mathbb{R}^1$ with the property that for $l = 1, 2, 3, \cdots$, there exists an $\varepsilon_l > 0$ and an $r_l > 0$ such that

$$|O_l(x)/x| \leqq r_l \quad \text{for all } |x| < \varepsilon_l, \qquad \varepsilon_l > 0, \quad r_l > 0. \tag{3.46}$$

3.47. LEMMA. *There exists an integer N and a function $O_1 : \mathbb{R}^1 \to \mathbb{R}^1$ (satisfying (3.46) for $l = 1$) such that for all $i \geqq N$ and $j = 0, 1, 2, \cdots, n-1$,*

$$\|H(z_{i+j}) - H(z_i)\| \leqq O_1(\|h_i\|), \tag{3.48}$$

where, as before, $H(z) \triangleq \partial^2 f(z)/\partial z^2$.

Proof. First, note that for $j \in \{0, 1, 2, \cdots, n-1\}$,

$$\|H(z_{i+j}) - H(z_i)\| \leqq \sum_{k=0}^{j-1} \|H(z_{i+k+1}) - H(z_{i+k})\|. \tag{3.49}$$

Next, since $f(\cdot)$ is three times continuously differentiable, we obtain from the Taylor formula,

$$\|H(z_{i+k+1}) - H(z_{i+k})\| \leqq \int_0^1 \|DH(z_{i+k} + t\lambda_{i+k}h_{i+k})(\lambda_{i+k}h_{i+k})\|\, dt, \tag{3.50}$$

where $DH(\cdot)(\cdot)$ denotes the third derivative of $f(\cdot)$. Since $f(\cdot)$ is three times continuously differentiable, and since $z_i \to \hat{z}$ as $i \to \infty$, there exists a $b \in (0, \infty)$ such that for $i + k = 0, 1, 2, \cdots$,

$$\|DH(z_{i+k} + t\lambda_{i+k}h_{i+k})\| \leqq b \quad \text{for all } t \in [0, 1]. \tag{3.51}$$

Consequently, because of (3.50), (3.29) and (3.42),

$$\|H(z_{i+j}) - H(z_i)\| \leqq nbq_4q_6\|h_i\| \triangleq O_1(\|h_i\|) \quad \text{for all } i \geqq N,$$

where N is such that (3.29) and (3.42) hold.

3.52. LEMMA. *There exists an integer N and a function $O_2: \mathbb{R}^1 \to \mathbb{R}^1$ (satisfying (3.46) for $l = 2$) such that for all $i \geqq N$ and $j = 0, 1, 2, \cdots, n-1$,*

$$\|H_{i+j} - H(z_i)\| \leqq O_2(\|h_i\|), \tag{3.53}$$

where H_i was defined in (2.26).

Proof. First, making use of (3.48) for $j = 0, 1, 2, \cdots, n-1$ and $i \geqq N$, we obtain

$$\begin{aligned} \|H_{i+j} - H(z_i)\| &\leqq \|H_{i+j} - H(z_{i+j})\| + \|H(z_{i+j}) - H(z_i)\| \\ &= \|H_{i+j} - H(z_{i+j})\| + O_1(\|h_i\|). \end{aligned} \tag{3.54}$$

Next, making use of (2.6) and the mean value theorem, we obtain

$$\begin{aligned} \|H_{i+j} - H(z_{i+j})\| &\leqq \int_0^1 \|H(z_{i+j} + t\lambda_{i+j}h_{i+j}) - H(z_{i+j})\|\, dt \\ &= \|H(z_{i+j} + \xi\lambda_{i+j}h_{i+j}) - H(z_{i+j})\| \end{aligned} \tag{3.55}$$

with $\xi \in [0, 1]$. Then, proceeding as in the proof of Lemma 3.47 we conclude that

$$\|H_{i+j} - H(z_i)\| \leqq O_1(\|h_i\|) + nbq_4q_6\|h_i\| \triangleq O_2(\|h_i\|) \quad \text{for } i \geqq N, \tag{3.56}$$

where N is such that (3.48) holds.

3.57. LEMMA. *There exists an integer N and functions $O_3(\cdot)$, $O_4(\cdot)$ and $O_5(\cdot)$, from $\mathbb{R}^1$ into $\mathbb{R}^1$ and satisfying* (3.46), *such that for $i \geqq N$, $i \in I_v$, and $j = 0, 1, 2, \cdots, n-1$,*

$$\|h_{i+j+1} - \overset{j+1}{h}_i\| \leqq O_3(\|h_{i+j} - \overset{j}{h}_i\|) + O_4(\|g_{i+j+1} - \overset{j+1}{g}_i\|) + O_5(\|h_i\|^2). \tag{3.58}$$

Proof. In what follows, we assume that $i \geqq N$, where N is an integer sufficiently large for all lemmas used to apply. First, suppose that $v = n$ and $j = n - 1$. Then, since $\omega(i + n) = 0$ for $i \in I_n$, it follows from (3.7) and (3.8) that

$$h_{i+n} = g_{i+n}, \qquad i \in I_n, \tag{3.59}$$

and hence, since $\overset{n}{h}_i = \overset{n}{g}_i = 0$, for $i \in I_n$,

$$\|h_{i+n} - \overset{n}{h}_i\| = \|g_{i+n} - \overset{n}{g}_i\|, \qquad i \in I_n. \tag{3.60}$$

Consequently, (3.58) is satisfied for $i \in I_v$ and $j = n - 1$ when $v = n$. In what follows, we assume that either $v > n$ and $j \in \{1, 2, \cdots, n-1\}$ or that $v = n$ and $j \in \{1, 2, \cdots, n-2\}$. Now,

$$\begin{aligned} \|h_{i+j+1} - \overset{j+1}{h}_i\| &= \|g_{i+j+1} + \gamma_{i+j}h_{i+j} - \overset{j+1}{g}_i - \overset{j}{\gamma}_i\overset{j}{h}_i\| \\ &\leqq \|g_{i+j+1} - \overset{j+1}{g}_i\| + \|\gamma_{i+j}h_{i+j} - \overset{j}{\gamma}_i\overset{j}{h}_i\|. \end{aligned} \tag{3.61}$$

We shall now obtain a bound on $\|\gamma_{i+j}h_{i+j} - \overset{j}{\gamma}_i\overset{j}{h}_i\|$ in (3.61). Thus, by (3.7) and (2.30),

$$\begin{aligned} \gamma_{i+j} &= \frac{\langle g_{i+j+1} - g_{i+j}, g_{i+j+1}\rangle}{\|g_{i+j}\|^2} \\ &= -\frac{\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle}{\|g_{i+j}\|^2}\lambda_{i+j} \\ &= -\frac{\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle}{\|g_{i+j}\|^2}\frac{\langle g_{i+j} - g_{i+j+1}, h_{i+j}\rangle}{\langle h_{i+j}, H_{i+j}h_{i+j}\rangle} \\ &= -\frac{\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle}{\langle h_{i+j}, H_{i+j}h_{i+j}\rangle} \\ &\quad -\gamma_{i+j-1}\frac{\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle\langle g_{i+j}, h_{i+j-1}\rangle}{\|g_{i+j}\|^2\langle h_{i+j}, H_{i+j}h_{i+j}\rangle} \\ &\quad +\frac{\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle\langle g_{i+j+1}, h_{i+j}\rangle}{\|g_{i+j}\|^2\langle h_{i+j}, H_{i+j}h_{i+j}\rangle}. \end{aligned} \tag{3.62}$$

From (3.34),

$$\overset{j}{\gamma}_i = -\frac{\langle \overset{j+1}{g}_i, H(z_i)\overset{j}{h}_i\rangle}{\langle \overset{j}{h}_i, H(z_i)\overset{j}{h}_i\rangle}. \tag{3.63}$$

Consequently,

$$\begin{aligned} \|\gamma_{i+j}h_{i+j} - \overset{j}{\gamma}_i\overset{j}{h}_i\| \leqq &\left\| -\frac{\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle}{\langle h_{i+j}, H_{i+j}h_{i+j}\rangle}h_{i+j} + \frac{\langle \overset{j+1}{g}_i, H(z_i)\overset{j}{h}_i\rangle}{\langle \overset{j}{h}_i, H(z_i)\overset{j}{h}_i\rangle}\overset{j}{h}_i \right\| \\ &+\left\| \frac{\gamma_{i+j-1}\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle\langle g_{i+j}, h_{i+j-1}\rangle}{\|g_{i+j}\|^2\langle h_{i+j}, H_{i+j}h_{i+j}\rangle}h_{i+j} \right\| \\ &+\left\| \frac{\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle\langle g_{i+j+1}, h_{i+j}\rangle}{\|g_{i+j}\|^2\langle h_{i+j}, H_{i+j}h_{i+j}\rangle}h_{i+j} \right\|. \end{aligned} \tag{3.64}$$

Next, since by construction $|\langle g_{i+j}, h_{i+j-1}\rangle| \leqq \|g_{i+j-1}\|\,\|g_{i+j}\|\,\|h_{i+j-1}\|$, and

making use of (1.9), (3.22), (3.25), (3.30) and (3.42), we obtain (see second term in (3.64))

$$
\begin{aligned}
p_{i+j} &\triangleq \frac{|\gamma_{i+j-1}|\,|\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle|\,|\langle g_{i+j}, h_{i+j-1}\rangle|\,\|h_{i+j}\|}{\|g_{i+j}\|^2\langle h_{i+j}, H_{i+j}h_{i+j}\rangle} \\
&\leqq \frac{q_1\|g_{i+j+1}\|M\|h_{i+j}\|\,\|g_{i+j}\|\,\|h_{i+j-1}\|^2\|h_{i+j}\|}{q_3\|g_{i+j}\|^2\|h_{i+j}\|^2 m} \qquad (3.65) \\
&\leqq \frac{q_1 q_6^2 M}{q_3 m}\,\frac{\|g_{i+j+1}\|}{\|g_{i+j}\|}\,\|h_i\|^2 \leqq \frac{q_1 q_2 q_5 q_6^2 M\|h_i\|^2}{q_3 m}.
\end{aligned}
$$

Similarly (see third term in (3.64)),

$$
\begin{aligned}
s_{i+j} &\triangleq \frac{|\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle|\,|\langle g_{i+j+1}, h_{i+j}\rangle|\,\|h_{i+j}\|}{\|g_{i+j}\|^2\langle h_{i+j}, H_{i+j}h_{i+j}\rangle} \\
&\leqq \frac{M\|g_{i+j+1}\|\,\|h_{i+j}\|\,\|g_{i+j}\|\,\|g_{i+j+1}\|\,\|h_{i+j}\|^2}{\|g_{i+j}\|^2 m\|h_{i+j}\|^2} \qquad (3.66) \\
&\leqq \frac{q_2 q_6^2 M\|h_i\|^2}{m}.
\end{aligned}
$$

Now, let

$$
t_{i+j} \triangleq \langle h_{i+j}, H_{i+j}h_{i+j}\rangle\langle \overset{j}{h}_i, H(z_i)\overset{j}{h}_i\rangle. \qquad (3.67)
$$

Then (see first term in (3.64)), by adding and subtracting terms, we obtain

$$
\begin{aligned}
v_{i+j} &\triangleq \left\| \frac{\langle \overset{j+1}{g}_i, H(z_i)\overset{j}{h}_i\rangle}{\langle \overset{j}{h}_i, H(z_i)\overset{j}{h}_i\rangle}\overset{j}{h}_i - \frac{\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle}{\langle h_{i+j}, H_{i+j}h_{i+j}\rangle}h_{i+j} \right\| \\
&= \frac{1}{t_{i+j}}\|\langle h_{i+j}, H_{i+j}h_{i+j}\rangle\langle \overset{j+1}{g}_i, H(z_i)\overset{j}{h}_i\rangle\overset{j}{h}_i \\
&\qquad\qquad - \langle \overset{j}{h}_i, H(z_i)\overset{j}{h}_i\rangle\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle h_{i+j}\| \\
&\leqq \frac{1}{t_{i+j}}\{\|\langle h_{i+j}, H_{i+j}h_{i+j}\rangle\langle \overset{j+1}{g}_i, H(z_i)(\overset{j}{h}_i - h_{i+j})\rangle\overset{j}{h}_i\| \\
&\qquad + \|\langle h_{i+j}, H_{i+j}(h_{i+j} - \overset{j}{h}_i)\rangle\langle \overset{j+1}{g}_i, H(z_i)h_{i+j}\rangle\overset{j}{h}_i\| \qquad (3.68) \\
&\qquad + \|\langle h_{i+j}, H_{i+j}\overset{j}{h}_i\rangle\langle \overset{j+1}{g}_i - g_{i+j+1}, H(z_i)h_{i+j}\rangle\overset{j}{h}_i\| \\
&\qquad + \|\langle h_{i+j} - \overset{j}{h}_i, H_{i+j}\overset{j}{h}_i\rangle\langle g_{i+j+1}, H(z_i)h_{i+j}\rangle\overset{j}{h}_i\| \\
&\qquad + \|\langle \overset{j}{h}_i, (H_{i+j} - H(z_i))\overset{j}{h}_i\rangle\langle g_{i+j+1}, H(z_i)h_{i+j}\rangle\overset{j}{h}_i\| \\
&\qquad + \|\langle \overset{j}{h}_i, H(z_i)\overset{j}{h}_i\rangle\langle g_{i+j+1}, (H(z_i) - H_{i+j})h_{i+j}\rangle\overset{j}{h}_i\| \\
&\qquad + \|\langle \overset{j}{h}_i, H(z_i)\overset{j}{h}_i\rangle\langle g_{i+j+1}, H_{i+j}h_{i+j}\rangle(\overset{j}{h}_i - h_{i+j})\|\}.
\end{aligned}
$$

Since, by (1.9),

$$\frac{1}{t_{i+j}} \leqq \frac{1}{m^2 \|h_{i+j}\|^2 \|\overset{j}{h}_i\|^2}, \tag{3.69}$$

and because of (1.9) and (3.56), we obtain from (3.67) and (3.68) that

$$\begin{aligned} v_{i+j} \leqq \frac{1}{m^2}\Bigg\{ & \frac{M^2 \|\overset{j+1}{g}_i\| \, \|\overset{j}{h}_i - h_{i+j}\|}{\|\overset{j}{h}_i\|} \\ & \frac{M^2 \|g_{i+j1+}\| \, \|\overset{j}{h}_i - h_{i+j}\|}{\|h_{i+j}\|} + \frac{M\|g_{i+j+1}\| O_2(\|h_i\|)\|\overset{j}{h}_i\|}{\|h_{i+j}\|} \\ & + \frac{M^2 \|g_{i+j+1}\| \, \|\overset{j}{h}_i - h_{i+j}\|}{\|h_{i+j}\|} + \frac{M\|g_{i+j+1}\| O_2(\|h_i\|)\|\overset{j}{h}_i\|}{\|h_{i+j}\|} \\ & + \frac{M\|g_{i+j+1}\| O_2(\|h_i\|)\|\overset{j}{h}_i\|}{\|h_{i+j}\|} + \frac{M^2 \|g_{i+j+1}\| \, \|\overset{j}{h}_i - h_{i+j}\|}{\|h_{i+j}\|} \Bigg\}. \end{aligned} \tag{3.70}$$

Finally, making use of (3.37), (3.30) and (3.44), we obtain,

$$\begin{aligned} v_{i+j} & \leqq \frac{M}{m^2}\{2\bar{q}_5 M \|\overset{j}{h}_i - h_{i+j}\| + M\|\overset{j+1}{g}_i - g_{i+j+1}\| \\ & \qquad + 2Mq_5 \|\overset{j}{h}_i - h_{i+j}\| + 2q_5 q_6 O_2(\|h_i\|)\|h_i\|\} \\ & = \frac{2M^2}{m^2}[\bar{q}_5 + q_5]\|\overset{j}{h}_i - h_{i+j}\| + \frac{M^2}{m^2}\|\overset{j+1}{g}_i - g_{i+j+1}\| \\ & \qquad + \frac{2M}{m^2} q_5 q_6 O_2(\|h_i\|)\|h_i\|. \end{aligned} \tag{3.71}$$

Now, from (3.61), (3.62), (3.64), (3.65), (3.66) and (3.68), we obtain

$$\begin{aligned} \|h_{i+j+1} - \overset{j+1}{h}_i\| & \leqq \|g_{i+j+1} - \overset{j+1}{g}_i\| + \|\overset{j}{\gamma}_i \overset{j}{h}_i - \gamma_{i+j} h_{i+j}\| \\ & \leqq \|g_{i+j+1} - \overset{j+1}{g}_i\| + v_{i+j} + p_{i+j} + s_{i+j} \\ & \leqq O_3(\|h_{i+j} - \overset{j}{h}_i\|) + O_4(\|g_{i+j+1} - \overset{j+1}{g}_i\|) + O_5(\|h_i\|^2), \end{aligned} \tag{3.72}$$

where O_3, O_4 and O_5 are defined in an obvious way from the relations (3.65), (3.66) and (3.71). Hence the lemma is true for $i \in I_v$ and $j = 1, 2, \cdots, n-1$.

Finally, for $i \in I_v$ and $j = 0$, $h_{i+j} = g_{i+j}$. Consequently (see (3.62)) for $i \in I_v$,

$$\gamma_i = -\frac{\langle g_{i+1}, H_i h_i\rangle}{\langle h_i, H_i h_i\rangle} + \frac{\langle g_{i+1}, H_i h_i\rangle \langle g_{i+1}, h_i\rangle}{\|g_i\|^2 \langle h_i, H_i h_i\rangle}.$$

As a result, proceeding as before we can show that $\|h_{i+1} - \overset{1}{h}_i\| \leqq \|g_{i+1} - \overset{1}{g}_i\| + v_i + s_i$, where v_i and s_i are defined as in (3.66) and (3.68) respectively, for $i = 0$.

Since $p_i \geqq 0$, it follows that (3.72) is also true for $j = 0$.

3.73. LEMMA. *There exists an integer N and a function $O_6: \mathbb{R}^1 \to \mathbb{R}^1$, satisfying* (3.46), *such that for all $i \geqq N$, $i \in I_v$, $j = 0, 1, \cdots, n-1$,*

$$\|g_{i+j+1} - \overset{j+1}{g_i}\| \leqq \|g_{i+j} - \overset{j}{g_i}\| + O_6(\|h_i\|^2) + M\|\lambda_{i+j}h_{i+j} - \overset{j}{\lambda_i}\overset{j}{h_i}\|. \tag{3.74}$$

Proof. Since by (2.30), $-g_{i+j+1} = -g_{i+j} + \lambda_{i+j}H_{i+j}h_{i+j}$ and, similarly, $-\overset{j+1}{g_i} = -\overset{j}{g_i} + \overset{j}{\lambda_i}H(\overset{j}{z_i})\overset{j}{h_i}$, it follows that for $i \in I_v$ and $j \in \{0, 1, 2, \cdots, n-1\}$,

$$\begin{aligned}
\|g_{i+j+1} - \overset{j+1}{g_i}\| &\leqq \|g_{i+j} - \overset{j}{g_i}\| + \|H_{i+j}\lambda_{i+j}h_{i+j} - H(\overset{j}{z_i})\overset{j}{\lambda_i}\overset{j}{h_i}\| \\
&\leqq \|g_{i+j} - \overset{j}{g_i}\| + \|(H_{i+j} - H(\overset{j}{z_i}))\overset{j}{\lambda_i}\overset{j}{h_i}\| \\
&\qquad + \|H_{i+j}(\lambda_{i+j}h_{i+j} - \overset{j}{\lambda_i}\overset{j}{h_i})\| \\
&\leqq \|g_{i+j} - \overset{j}{g_i}\| + \bar{q}_4 q_6 O_2(\|h_i\|)\|h_i\| \\
&\qquad + M\|\lambda_{i+j}h_{i+j} - \overset{j}{\lambda_i}\overset{j}{h_i}\|,
\end{aligned} \tag{3.75}$$

where we have made use of (1.9), (3.38), (3.44) and (3.53). Setting $O_6(\|h_i\|^2) \equiv \bar{q}_4 q_6 O_2(\|h_i\|)\|h_i\|$, we obtain (3.74) from (3.75).

3.76. LEMMA. *There exists an integer N and functions $O_7(\cdot)$, $O_8(\cdot)$ and $O_9(\cdot)$, from $\mathbb{R}^1$ into $\mathbb{R}^1$ and satisfying* (3.46), *such that for all $i \geqq N$, $i \in I_v$, and $j = 0$, $1, \cdots, n-1$,*

$$\|\lambda_{i+j}h_{i+j} - \overset{j}{\lambda_i}\overset{j}{h_i}\| \leqq O_7(\|g_{i+j} - \overset{j}{g_i}\|) + O_8(\|h_{i+j} - \overset{j}{h_i}\|) + O_9(\|h_i\|^2). \tag{3.77}$$

Proof. Suppose that $i \in I_v$, $j \in \{0, 1, \cdots, n-1\}$. Making use of (2.32) and (3.33) we obtain

$$\begin{aligned}
\|\lambda_{i+j}h_{i+j} - \overset{j}{\lambda_i}\overset{j}{h_i}\| \leqq &\left\| \frac{\langle g_{i+j}, h_{i+j}\rangle}{\langle h_{i+j}, H_{i+j}h_{i+j}\rangle}h_{i+j} - \frac{\langle \overset{j}{g_i}, \overset{j}{h_i}\rangle}{\langle \overset{j}{h_i}, H(\overset{j}{z_i})\overset{j}{h_i}\rangle}\overset{j}{h_i} \right\| \\
&+ \left\| \frac{\langle g_{i+j+1}, h_{i+j}\rangle}{\langle h_{i+j}, H_{i+j}h_{i+j}\rangle}h_{i+j} \right\|.
\end{aligned} \tag{3.78}$$

Now, making use of (1.9), of the test in Step 13 of Algorithm 3.1, (3.41) and (3.42), we obtain

$$\frac{|\langle g_{i+j+1}, h_{i+j}\rangle|\,\|h_{i+j}\|}{\langle h_{i+j}, H_{i+j}h_{i+j}\rangle} \leqq \frac{\|g_{i+j}\|\,\|g_{i+j+1}\|\,\|h_{i+j}\|^2}{m\|h_{i+j}\|^2} \leqq \frac{q_6^2\|h_i\|^2}{m}. \tag{3.79}$$

Next, defining t_{i+j} as in (3.67), and first adding and subtracting terms and then making use of (1.9), (3.25), (3.36) and (3.35), we obtain

$$\left\| \frac{\langle g_{i+j}, h_{i+j}\rangle}{\langle h_{i+j}, H_{i+j}h_{i+j}\rangle}h_{i+j} - \frac{\langle \overset{j}{g_i}, \overset{j}{h_i}\rangle}{\langle \overset{j}{h_i}, H(\overset{j}{z_i})\overset{j}{h_i}\rangle}\overset{j}{h_i} \right\|$$

$$= \frac{1}{t_{i+j}}\|\langle g_{i+j}, h_{i+j}\rangle\langle \overset{j}{h_i}, H(\overset{j}{z_i})\overset{j}{h_i}\rangle h_{i+j} - \langle \overset{j}{g_i}, \overset{j}{h_i}\rangle\langle h_{i+j}, H_{i+j}h_{i+j}\rangle\overset{j}{h_i}\|$$

$$
\begin{aligned}
&\leqq \frac{1}{t_{i+j}}\{\|\langle g_{i+j}, h_{i+j} - \overset{j}{h}_i\rangle\langle \overset{j}{h}_i, H(z_i)\overset{j}{h}_i\rangle h_{i+j}\| \\
&\quad + \|\langle g_{i+j}, \overset{j}{h}_i\rangle\langle \overset{j}{h}_i - h_{i+j}, H(z_i)\overset{j}{h}_i\rangle h_{i+j}\| \\
&\quad + \|\langle g_{i+j} - \overset{j}{g}_i, \overset{j}{h}_i\rangle\langle h_{i+j}, H(z_i)\overset{j}{h}_i\rangle h_{i+j}\| \\
&\quad + \|\langle \overset{j}{g}_i, \overset{j}{h}_i\rangle\langle h_{i+j}, H(z_i)(\overset{j}{h}_i - h_{i+j})\rangle h_{i+j}\| \\
&\quad + \|\langle \overset{j}{g}_i, \overset{j}{h}_i\rangle\langle h_{i+j}, (H(z_i) - H_{i+j})h_{i+j}\rangle h_{i+j}\| \\
&\quad + \|\langle \overset{j}{g}_i, \overset{j}{h}_i\rangle\langle h_{i+j}, H_{i+j}h_{i+j}\rangle(h_{i+j} - \overset{j}{h}_i)\|\} \\
&\leqq \frac{1}{m^2}\left\{\frac{M\|g_{i+j}\|\,\|h_{i+j} - \overset{j}{h}_i\|}{\|h_{i+j}\|} + \frac{M\|g_{i+j}\|\,\|h_{i+j} - \overset{j}{h}_i\|}{\|h_{i+j}\|}\right. \\
&\quad + M\|g_{i+j} - \overset{j}{g}_i\| + \frac{M\|\overset{j}{g}_i\|\,\|h_{i+j} - \overset{j}{h}_i\|}{\|\overset{j}{h}_i\|} \\
&\quad \left. + \frac{\|\overset{j}{g}_i\|\,\|h_{i+j}\|O_2(\|h_i\|)}{\|\overset{j}{h}_i\|} + \frac{M\|\overset{j}{g}_i\|\,\|h_{i+j} - \overset{j}{h}_i\|}{\|\overset{j}{h}_i\|}\right\} \\
&\leqq \frac{1}{m^2}\left\{2M\left(\frac{1}{q_3} + \frac{1}{\bar{q}_3}\right)\|h_{i+j} - \overset{j}{h}_i\| + M\|g_{i+j} - \overset{j}{g}_i\|\right. \\
&\quad \left. + \frac{q_6}{\bar{q}_3}O_2(\|h_i\|)\|h_i\|\right\}.
\end{aligned} \tag{3.80}
$$

Since the existence of functions $O_7(\cdot)$, $O_8(\cdot)$ and $O_9(\cdot)$ satisfying (3.77) follows directly from (3.78), (3.79) and (3.80), we are done.

3.81. LEMMA. *There exists an integer N and functions $O_{10}(\cdot)$, $O_{11}(\cdot)$ and $O_{12}(\cdot)$ from $\mathbb{R}^1$ into $\mathbb{R}^1$ and satisfying* (3.46) *such that for $i \geqq N$, $i \in I_v$, and $j = 0, 1, 2, \cdots, n-1$,*

$$\|g_{i+j} - \overset{j}{g}_i\| \leqq O_{10}(\|h_i\|^2), \tag{3.82}$$

$$\|h_{i+j} - \overset{j}{h}_i\| \leqq O_{11}(\|h_i\|^2), \tag{3.83}$$

$$\|\lambda_{i+j}h_{i+j} - \overset{j}{\lambda}_i\overset{j}{h}_i\| \leqq O_{12}(\|h_i\|^2). \tag{3.84}$$

Proof. We make use of induction. Let N be an integer for which the conclusions of Lemmas 3.73, 3.57 and 3.76 hold, and let i be any positive integer satisfying $i \geqq N$ and $i \in I_v$. Then, for $j = 0$, $\overset{0}{g}_i = g_i = \overset{0}{h}_i = h_i$, and hence

$$\|g_i - \overset{0}{g}_i\| = \|h_i - \overset{0}{h}_i\| = 0 \tag{3.85}$$

and

$$\|\lambda_i h_i - \overset{0}{\lambda}_i \overset{0}{h}_i\| \leqq O_9(\|h_i\|^2). \tag{3.86}$$

Now, for any $j \in \{0, 1, 2, \cdots, n-2\}$, suppose that there exist functions $\overset{j}{O}_{10}(\cdot)$, $\overset{j}{O}_{11}(\cdot)$ and $\overset{j}{O}_{12}(\cdot)$, satisfying (3.46), such that

$$\|g_{i+j} - \overset{j}{g}_i\| \leqq \overset{j}{O}_{10}(\|h_i\|^2), \tag{3.87}$$

$$\|h_{i+j} - \overset{j}{h}_i\| \leqq \overset{j}{O}_{11}(\|h_i\|^2), \tag{3.88}$$

$$\|\lambda_{i+j} h_{i+j} - \overset{j}{\lambda}_i \overset{j}{h}_i\| \leqq \overset{j}{O}_{12}(\|h_i\|^2). \tag{3.89}$$

Then, from Lemma 3.73, (3.87) and (3.89),

$$\begin{aligned} \|g_{i+j+1} - \overset{j+1}{g}_i\| &\leqq \overset{j}{O}_{10}(\|h_i\|^2) + O_6(\|h_i\|^2) + M\overset{j}{O}_{12}(\|h_i\|^2) \\ &\triangleq \overset{j+1}{O}_{10}(\|h_i\|^2), \end{aligned} \tag{3.90}$$

where, obviously, $\overset{j+1}{O}_{10}(\cdot)$ is a function satisfying (3.46).

Next, from (3.58), (3.88) and (3.90),

$$\begin{aligned} \|h_{i+j+1} - \overset{j+1}{h}_i\| \leqq{}& \frac{O_3(\|h_{i+j} - \overset{j}{h}_i\|)}{\|h_{i+j} - \overset{j}{h}_i\|} \overset{j}{O}_{11}(\|h_i\|^2) + \frac{O_4(\|g_{i+j+1} - \overset{j+1}{g}_i\|)}{\|g_{i+j+1} - \overset{j+1}{g}_i\|} \overset{j}{O}_{10}(\|h_i\|^2) \\ &+ O_5(\|h_i\|^2). \end{aligned} \tag{3.91}$$

Since (3.88) and (3.90) hold, it follows from (3.91) that there exists a function $\overset{j+1}{O}_{11}(\cdot)$ satisfying (3.46) such that

$$\|h_{i+j+1} - \overset{j+1}{h}_i\| \leqq \overset{j+1}{O}_{11}(\|h_i\|^2). \tag{3.92}$$

Finally, making use of (3.77), (3.90) and (3.92) we obtain

$$\begin{aligned} \|\lambda_{i+j+1} h_{i+j+1} - \overset{j+1}{\lambda}_i \overset{j+1}{h}_i\| \leqq{}& \frac{O_7(\|g_{i+j+1} - \overset{j+1}{g}_i\|)}{\|g_{i+j+1} - \overset{j+1}{g}_i\|} \overset{j+1}{O}_{10}(\|h_i\|^2) \\ &+ \frac{O_8(\|h_{i+j+1} - \overset{j+1}{h}_i\|)}{\|h_{i+j+1} - \overset{j+1}{h}_i\|} \overset{j+1}{O}_{11}(\|h_i\|^2) + O_9(\|h_i\|^2). \end{aligned} \tag{3.93}$$

Since (3.90) and (3.92) are satisfied, (3.93) implies that there exists a function $\overset{j+1}{O}_{12}(\cdot)$ satisfying (3.46) such that

$$\|\lambda_{i+j+1} h_{i+j+1} - \overset{j+1}{\lambda}_i \overset{j+1}{h}_i\| \leqq \overset{j+1}{O}_{12}(\|h_i\|^2). \tag{3.94}$$

Since (3.87)–(3.89) are true for $j = 0$, we see that they must also be true for $j = 1$,

$2, \cdots, n-1$. To complete the proof, we set $O_{10}(\cdot) = \max_j \overset{j}{O}_{10}(\cdot)$, $O_{11}(\cdot) = \max_j \overset{j}{O}_{11}(\cdot)$ and $O_{12}(\cdot) = \max_j \overset{j}{O}_{12}(\cdot)$.

We are finally ready to establish our main result.

3.95. THEOREM. *There exists an integer N' and a $q \in (0, \infty)$ such that for all $i \geqq N'$, $i \in I_v$, and $j = 0, 1, 2, \cdots, n-1$,*

$$\|\lambda_{i+j} h_{i+j} - \overset{j}{\lambda}_i \overset{j}{h}_i\| \leqq q \|z_i - \hat{z}\|^2, \tag{3.96}$$

where $\hat{z}$ is the unique minimizer of $f(\cdot)$.

Proof. First, by the Taylor formula,

$$-g_i = \nabla f(z_i) = \nabla f(\hat{z}) + \int_0^1 H(z_i + t(z_i - \hat{z}))\, dt (z_i - \hat{z}), \tag{3.97}$$

and hence, because of (1.9) and because $\nabla f(\hat{z}) = 0$,

$$\|g_i\| \leqq M \|z_i - \hat{z}\|. \tag{3.98}$$

Now, since $g_i \to 0$ as $i \to \infty$, because $z_i \to \hat{z}$ as $i \to \infty$, it follows from (3.25) that $h_i \to 0$ as $i \to \infty$. Consequently, because of (3.84), (3.46), (3.25) and (3.98), there exists an integer $N' \geqq N$ such that for all $i \geqq N'$, $i \in I_v$, and $j = 0, 1, 2, \cdots, n-1$,

$$\begin{aligned} \|\lambda_{i+j} h_{i+j} - \overset{j}{\lambda}_i \overset{j}{h}_i\| &\leqq O_{12}(\|h_i\|^2) \leqq r_{12} \|h_i\|^2 \\ &\leqq r_{12} q_2^2 \|g_i\|^2 \leqq r_{12} q_2^2 M^2 \|z_i - \hat{z}\|^2 \\ &\triangleq q \|z_i - \hat{z}\|^2, \end{aligned} \tag{3.99}$$

where $r_{12} > 0$ is such that (3.46) holds for $l = 12$. This completes our proof.

Thus, the assumptions of Theorem 3.13 are indeed satisfied.

Conclusion. We have shown in this paper that it is possible to construct a superlinearly convergent, conjugate gradient algorithm which does not include the minimization of a function along a line as a subprocedure. A most important consequence of this is that unlike some of its "theoretical" predecessors, our algorithm is directly implementable; i.e. it can be programmed as stated, without any need for heuristics to circumvent nonimplementable operations. Of the two versions stated in this paper, we have programmed the first one, Algorithm 2.1, and have tested it against a few standard problems such as the Rosenbrock's valley. The empirical results show that this version converges at about the same rate as the more complex version Algorithm 3.1 (i.e., superlinearly), and hence, this is the version which we would normally recommend.

Appendix. Properties of the linear search subprocedure. Suppose $\theta: \mathbb{R}^1 \to \mathbb{R}^1$ is twice continuously differentiable and satisfies

$$0 < \overline{m} \leqq \theta''(x) \leqq \overline{M} < \infty \tag{A.1}$$

for all $x \in \mathbb{R}^1$. The following algorithm will be shown to yield a sequence converging to the minimum of $\theta(\cdot)$.

A.2. LINEAR SEARCH PROCEDURE.

Step 0. Set $l = 0$ and choose any $x_0 \in \mathbb{R}^1$ and any $\beta \in (0, 1)$.

Step 1. Compute $\theta'(x_l)$.

Step 2. If $\theta'(x_l) = 0$, stop; else set

$$\eta(x_l) = \begin{cases} \theta'(x_l), & l = 0, \\ [(x_l - x_{l-1})/(\theta'(x_l) - \theta'(x_{l-1}))]\theta'(x_l), & l \geqq 1, \end{cases} \tag{A.3}$$

and go to Step 3.

Step 3. Compute the smallest nonnegative integer $j(x_l)$ such that

$$\theta(x_l - \beta^{j(x_l)}\eta(x_l)) - \theta(x_l) + [\beta^{j(x_l)}/3]\eta(x_l)\theta'(x_l) \leqq 0. \tag{A.4}$$

Step 4. Set $x_{l+1} = x_l - \beta^{j(x_l)}\eta(x_l)$.

Step 5. Set $l = l + 1$ and go to Step 1.

A.5. THEOREM. *Suppose the linear search procedure* A.2 *generates an infinite sequence,* $\{x_l\}_{l=0}^{\infty}$. *Then* $x_l \to \hat{x}$ *as* $l \to \infty$, *where* $\hat{x}$ *is the minimizer of* $\theta(\cdot)$.

Proof. First, by use of the mean value theorem and (A.1), we conclude that

$$\eta(x_l)\theta'(x_l) \geqq \theta'(x_l)^2/\overline{M}, \qquad l = 1, 2, \cdots. \tag{A.6}$$

Also by use of the mean value theorem,

$$\begin{aligned} &\theta(x_l - \lambda\eta(x_l)) - \theta(x_l) + \frac{\lambda}{3}\eta(x_l)\theta'(x_l) \\ &\quad = -\theta'(x_l)\lambda\eta(x_l) + \frac{\lambda^2}{2}\theta''(\xi_l)\eta(x_l)^2 + \frac{\lambda}{3}\eta(x_l)\theta'(x_l) \\ &\quad = -\frac{2}{3}\lambda\frac{\theta'(x_l)^2}{\theta''(\zeta_l)} + \frac{1}{2}\lambda^2\theta''(\xi_l)\frac{\theta'(x_l)^2}{\theta''(\zeta_l)^2} \\ &\quad = \lambda\frac{\theta'(x_l)^2}{\theta''(\zeta_l)}\left[-\frac{2}{3} + \frac{\lambda}{2}\frac{\theta''(\xi_l)}{\theta''(\zeta_l)}\right], \end{aligned} \tag{A.7}$$

where $\xi_l \in [x_l, x_l - \lambda\eta(x_l)]$, $\zeta_l \in [x_{l-1}, x_l]$ and $l \geqq 1$. Hence, for $\lambda \geqq 0$ and $l \geqq 1$, we can use (A.1) in (A.7) to find

$$\theta(x_l - \lambda\eta(x_l)) - \theta(x_l) + \frac{\lambda}{3}\eta(x_l)\theta'(x_l) \leqq \lambda\frac{\theta'(x_l)^2}{\theta''(\zeta_l)}\left[-\frac{2}{3} + \frac{\lambda}{2}\frac{\overline{M}}{\overline{m}}\right]. \tag{A.8}$$

Consequently, we conclude that $j(x_l) \leqq \bar{j}$, where $\bar{j}$ is the smallest integer such that $-\frac{2}{3} + \frac{1}{2}(\overline{M}\beta^{\bar{j}}/\overline{m}) \leqq 0$. Thus, from (A.4) and (A.6),

$$\theta(x_{l+1}) - \theta(x_l) \leqq -\frac{\beta^{j(x_l)}}{3}\eta(x_l)\theta'(x_l) \leqq \frac{-\beta^{\bar{j}}}{3\overline{M}}\theta'(x_l)^2 \tag{A.9}$$

for $l = 1, 2, \cdots$.

Now, (A.1) implies that $\{x|\theta(x) \leqq \theta(x_0)\}$ is compact, and hence $\{x_l\}$ must have accumulation points. Furthermore, if x^* is an accumulation point such that

$\theta'(x^*) \neq 0$ and K indexes a subsequence such that $x_l \to x^*$ as $l \to \infty$, $l \in K$, we can use the continuity of $\theta'(\cdot)$ and the monotonicity of $\{\theta(x_l)\}$ to find an integer k such that

$$\theta(x_{l+j}) - \theta(x_l) \leqq -\frac{\beta^{\bar{j}}}{6\bar{M}}\theta'(x^*) < 0 \tag{A.10}$$

for adjacent indices l and $l + j$ in K and all $l \geqq k$. But this implies that $\{\theta(x_l)\}_{l\in K}$ is not Cauchy, which is impossible since $\theta(x_l) \to \theta(x^*)$ as $l \to \infty$, $l \in K$. Thus, we must conclude that all accumulation points of $\{x_l\}$ are stationary for $\theta(\cdot)$. It then follows from (A.1) and the compactness of $\{x|\theta(x) \leqq \theta(x_0)\}$ that $x_l \to \hat{x}$ as $l \to \infty$, where $\hat{x}$ is the minimum of $\theta(\cdot)$.

A.11. LEMMA. *Suppose Algorithm* A.2 *generates an infinite sequence* $\{x_l\}_{l=0}^{\infty}$. *Then there exists an integer* k *such that* $j(x_l) = 0$ *for all* $l \geqq k$.

Proof. This result follows directly from (A.7) and the fact that Theorem A.5 implies that $[\theta''(\xi_l)/\theta''(\zeta_l)] \to 1$ as $l \to \infty$.

In view of Lemma A.11, we see that for all l sufficiently large,

$$x_{l+1} = x_l - \frac{x_l - x_{l-1}}{\theta'(x_l) - \theta'(x_{l-1})}\theta'(x_l) \tag{A.12}$$

which is just the well-known secant method for finding the root of $\theta'(\cdot)$. Also, since Theorem A.5 guarantees that $x_l \to \hat{x}$ as $l \to \infty$, where $\theta'(\hat{x}) = 0$, we can deduce the following result from [12, (11.2.8)].

A.13. THEOREM. *If Algorithm* A.2 *constructs an infinite sequence* $\{x_l\}_{l=0}^{\infty}$, *then* $x_l \to \hat{x}$ *as* $l \to \infty$ (*where* $\theta'(\hat{x}) = 0$) *superlinearly*; *i.e.*,

$$\limsup_{l\to\infty} \frac{|x_{l+1} - \hat{x}|}{|x_l - \hat{x}|} = 0. \tag{A.14}$$

REFERENCES

[1] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[2] E. POLAK AND G. RIBIÈRE, *Note sur la convergence de methodes de directions conjugées*, Revue Francaise Inf. Rech. Oper., 16 RI (1969), pp. 35–43.

[3] R. FLETCHER AND C. M. REEVES, *Function minimization by conjugate gradients*, British Comp. J., 7 (1964), pp. 149–154.

[4] A. COHEN, *Rate of convergence and optimality conditions for root finding and optimization algorithms*, Doctoral thesis, Univ. of California, Berkeley, 1970.

[5] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[6] J. DANIEL, *The conjugate gradient method for linear and nonlinear operator equations*, SIAM J. Numer. Anal., 4 (1967), pp. 10–26.

[7] E. POLAK, *On the implementation of conceptual algorithms*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1970.

[8] B. T. POLYAK, *Conjugate gradient method in extremal problems*, U.S.S.R. Comput. Math. and Math. Phys., 9 (1969), pp. 809–821.

[9] R. M. ELKINS, *Convergence theorems for Gauss–Seidel and other minimization algorithms*, Computer Science Center Tech. Rep. 68-59, Univ. of Maryland, College Park, 1968.

[10] J. DANIEL, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[11] M. HESTENES AND E. STEIFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[12] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[13] R. FLETCHER AND M. J. D. POWELL, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

NECESSARY CONDITIONS FOR CONTINUOUS PARAMETER STOCHASTIC OPTIMIZATION PROBLEMS*

H. J. KUSHNER†

Abstract. A maximum principle is derived for the problem where the system is the Itô equation

$$dx = f(x, u, t)\,dt + \sigma(x, t)\,dz, \qquad 0 \leqq t \leqq T,$$

and the cost is of the form $Ex_0(T) + Eh(x(T))$, where $x_0(T)$ is the zeroth component of $x(T)$, and T is a real number. There are constraints of the type $E\tilde{r}_0(x(0)) = 0$, $Er_i(x(t_i), Ex(t_i)) = 0$, $i = 1, \cdots, k$, $E\tilde{q}_i(x(t_i), Ex(t_i)) \leqq 0$, $i = 0, 1, \cdots, k$, where t_i are given real numbers. The paper adapts the general maximum principle of Neustadt to the above stochastic problem.

1. Introduction. This paper applies the abstract variational theory of Neustadt [1] to obtain a stochastic maximum principle. Since the papers of Kushner on the stochastic maximum principle [2], [3], a number of developments were reported in Brodeau [4], Baum [5], Fleming [6], Sworder [7]–[8]. The versatile mathematical programming ideas were not used explicitly in [2]–[8], and, with relative ease, we are able to handle greater varieties of state space constraints than treated in the references. A discrete parameter analogue of the discrete maximum principle of Halkin [9] and Holtzman [10] appears in Kushner [11]. Even in the deterministic case, the ability to handle general constraints with relative ease gives the programming approach a distinct advantage over more direct approaches.

It is premature to assert that the stochastic maximum principle will be useful in providing any deep understanding of stochastic control problems. Nevertheless, it seems likely that the implicit geometric framework (at least in the programming approach) will suggest some useful approximation or numerical procedures. The results may serve as a departure point for a perturbation analysis as in the formal work [12], and the nature and interpretation of the random multipliers may shed additional light on the physical interpretation of the derivatives (weak or strong) of the minimum cost function which appears in the dynamic programming formulation for a fully Markovian problem. These various points are under current investigation for both the present work and [11]. Even for an initially Markovian problem, dynamic programming is not always applicable when there are state space constraints, and the alternative programming formulation may be useful to shed light on the control problem. For a discussion, for an elementary stochastic control problem, of the relationship between randomized controls and "singular arcs" see [13].

The problem formulation and mathematical background are given in § 2. A required result of Neustadt is stated in § 3 and the linearized equations are discussed in § 4. Section 5 derives a certain convex cone. The maximum principle is stated in

* Received by the editors January 11, 1971, and in final revised form August 30, 1971.

† Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported in part by the National Science Foundation under Grant GK 2788, 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 693-67B.

§ 6. The development in §§ 4–6 is for the open loop case and extensions are discussed in § 7.

2. Problem formulation and mathematical background.

A remark on notation. Let $m(\cdot, \cdot)$ denote an arbitrary random function with values $m(\omega, t)$, $0 \leqq t \leqq T$. The notation will be simplified by omitting the ω variable. The term $m(t)$ will be used for both $m(\cdot, t)$ and $m(\omega, t)$, and either $m(\cdot)$ or m (depending on the context) will be used for the random function $m(\cdot, \cdot)$. A random variable $M(\cdot)$ with value $M(\omega)$ will be written simply as M.

R^n denotes an n-dimensional Euclidean space.

Assumptions. Let[1] $z(\cdot) = (z_0(\cdot), \cdots, z_n(\cdot))'$, $0 \leqq t \leqq T$, be an $(n+1)$-dimensional normalized Brownian motion on the probability triple $(\Omega, P(\cdot), \mathscr{B})$, where Ω is the sample space, and $P(\cdot)$ the measure on the σ-algebra $\mathscr{B}$ on Ω. For any finite-dimensional vector $\alpha = (\alpha_1, \cdots, \alpha_r)$ and matrix $\Phi = \{\varphi_{ij}, i,j = 1, \cdots, r\}$, define the Euclidean norms $|\alpha|^2 = \sum_i |\alpha_i|^2$, $|\Phi|^2 = \sum_{i,j} \varphi_{ij}^2$. The control is an n_1-dimensional random function whose properties are described in Assumption 2.1 below. Let $f(\cdot, \cdot, \cdot)$ denote an R^{n+1}- valued function on $R^{n+1} \times R^{n_1} \times [0, T]$ and $\sigma(\cdot, \cdot)$ an $(n+1) \times (n+1)$ matrix-valued function on $R^{n+1} \times [0, T]$. Further properties of $f(\cdot, \cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ are given in Assumption 2.2 below. The control system of concern is the $(n+1)$-dimensional stochastic differential equation (of the Itô type) (1) on the fixed time interval $[0, T]$:

$$\begin{aligned} dx(t) &= f(x(t), u(t), t)\, dt + \sigma(x(t), t)\, dz(t), \\ x(t) &= (x_0(t), \cdots, x_n(t))'. \end{aligned} \tag{1}$$

The control $u(\cdot)$ and $x(0)$ satisfy Assumption 2.1 below. Write $\sigma(\alpha, t) = [\sigma_0(\alpha, t), \cdots, \sigma_n(\alpha, t)]$, where $\sigma_i(\alpha, t)$ is the ith column of $\sigma(\alpha, t)$.

$$dx(t) = f(x(t), u(t), t)\, dt + \sum_i \sigma_i(x(t), t)\, dz_i(t). \tag{2}$$

Let $h(\cdot)$ be a real-valued Borel function on R^{n+1} for which $Eh(x(T))$ exists for the $x(T)$ corresponding to any admissible control (see Assumption 2.1 below). Let $t_0, \cdots, t_{k+1}$, for a fixed integer k, denote a sequence of fixed times satisfying $0 = t_0 < t_1 < \cdots < t_{k+1} = T$. Let $a_0, \cdots, a_{k+1}$ and b_0 and b_1 be given positive integers. Let $\tilde{q}_i^j(\cdot, \cdot)$, $i = 0, \cdots, k+1$, $j = 1, \cdots, a_i$, and $\tilde{r}_0^j(\cdot, \cdot)$, $j = 1, \cdots, b_0$, and $\tilde{r}_{k+1}^j(\cdot, \cdot)$, $j = 1, \cdots, b_1$, be real-valued Borel functions on $R^{n+1} \times R^{n+1}$ and define

$$\begin{aligned} \tilde{q}_i(\cdot, \cdot) &= (\tilde{q}_i^1(\cdot, \cdot), \cdots, \tilde{q}_i^{a_i}(\cdot, \cdot))', \\ \tilde{r}_0(\cdot, \cdot) &= (\tilde{r}_0^1(\cdot, \cdot), \cdots, \tilde{r}^{b_0}(\cdot, \cdot))', \\ \tilde{r}_{k+1}(\cdot, \cdot) &= (\tilde{r}_{k+1}^1(\cdot, \cdot), \cdots, \tilde{r}_{k+1}^{b_1}(\cdot, \cdot))'. \end{aligned}$$

For any admissible control (see Assumption 2.1), let the corresponding $x(t_0)$, $\cdots$, $x(t_{k+1})$ satisfy $E|\tilde{q}_i(x(t_i), Ex(t_i))| < \infty$, $i = 0, \cdots, k+1$, and $E|\tilde{r}_i(x(t_i), Ex(t_i))| < \infty$, $i = 0, k+1$ (properties guaranteed by Assumption 2.4 below).

The problem. Use Assumptions 2.1 and 2.2 and the above properties on h, $\tilde{q}_i$, $\tilde{r}_i$. Define the cost function

$$\varphi_0(x(\cdot)) \equiv Ex_0(T) + Eh(x(T)). \tag{3}$$

[1] The $'$ denotes transpose.

In the class of admissible controls for which the corresponding trajectories satisfy the constraints

$$
\begin{aligned}
q_i(x(t_i)) &\equiv E\tilde{q}_i(x(t_i), Ex(t_i)) \leqq 0, && i = 0, \cdots, k+1, \\
r_i(x(t_i)) &\equiv E\tilde{r}_i(x(t_i), Ex(t_i)) = 0, && i = 0, k+1,
\end{aligned} \tag{4}
$$

assume that there is one, denoted by $\hat{u}$, for which the cost is minimized (or is no greater for any other control in the class). It is assumed that $q_0(x(\cdot)) = 0$ implies that $x_0(0) = 0$. As discussed below, more general constraints can be treated. Let $\hat{x}$ denote the corresponding optimal solution to (1).

Now, add Assumptions 2.3–2.5 and find a necessary condition for $\hat{u}$ and $\hat{x}$.

Assumption 2.1. Let $\mathscr{B}_t$, $T \geqq t \geqq 0$, denote a family of given σ-algebras which are nonanticipative with respect to the Wiener $z(\cdot)$ process. The $\mathscr{B}_t$ are the data σ-algebras. $x(0)$ is measurable on $\mathscr{B}_0$ and $E|x(0)|^2 < \infty$. Let $\mathscr{U}_t$, $T \geqq t \geqq 0$, denote a family of given nonempty n_1-dimensional sets. The family of admissible controls, denoted by $\tilde{\mathscr{U}}$, is the collection of n_1-dimensional random functions $u(\cdot, \cdot)$, with values $u(\omega, t)$ in $\mathscr{U}_t$ at time t, where $u(\cdot, t)$ is measurable over $\mathscr{B}_t$. As noted above, we shall write either u or $u(\cdot)$ for the function $u(\cdot, \cdot)$, and $u(t)$ for either $u(\omega, t)$ or $u(\cdot, t)$.

Assumption 2.2. The $f(\cdot, \cdot, \cdot)$ and $\sigma_i(\cdot, \cdot)$ are Borel functions of their arguments. $f(\cdot, \beta, t)$ and $\sigma_i(\cdot, t)$ are differentiable for each fixed β, t and t, respectively. Write $f_x(\alpha, \beta, t)$ and $\sigma_{i,x}(\alpha, t)$ for the matrices with j,kth elements $\partial f_j(\alpha, \beta, t)/\partial \alpha_k$ and $\partial \sigma_{ij}(\alpha, t)/\partial \alpha_k$, respectively, and suppose that both are uniformly bounded. Assume $|f(\alpha, \beta, t)|^2 \leqq K_0(1 + |\alpha|^2)$, $|\sigma_i(\alpha, t)|^2 \leqq K_0(1 + |\alpha|^2)$, uniformly in $\mathscr{B} \in \mathscr{U}_t$ and $t \in [0, T]$. The function $f(\cdot, \beta, t)$ is continuous at each $\beta \in \mathscr{U}_t$ and $t \in [0, T]$, uniformly in t.

Assumption 2.3. For each fixed $t \in (0, T]$ and $\mathscr{B}_t$ measurable and $\mathscr{U}_t$-valued *random variable* u_t, there is a $\delta(t) > 0$ so that for each $\delta < \delta(t)$ there is a *random variable* $\tilde{u}_{t-\delta}$ with the property that $\tilde{u}_{t-\delta}$ is measurable over each $\mathscr{B}_s$ and has values in each $\mathscr{U}_s$, where $s \in [t - \delta, t]$, and the function $\tilde{u}_{t-\delta}$ satisfies

$$
f(\hat{x}(t), u_t, t) - f(\hat{x}(t), \tilde{u}_{t-\delta}, t) \to 0 \tag{5}
$$

in probability as $\delta \to 0$. Both $\tilde{u}_{t-\delta}$ and $\delta(t)$ may depend on u_t and t.

Note. The condition of the last paragraph is included since we shall use piecewise constant and nonanticipative perturbations to the optimal control. Its intuitive meaning is simply that the effect of any random control u_t which can be used at time t can be approximated by some random control $\tilde{u}_{t-\delta}$ which can be used at any time in the small interval $[t - \delta, t]$.

Assumption 2.4. Assume that, for any R^{n+1}-valued random variable v,

$$
\begin{aligned}
|q_i(v)| &\leqq K_0(1 + E|v|^2), && i = 0, 1, \cdots, k+1, \\
|r_i(v)| &\leqq K_0(1 + E|v|^2), && i = 0, k+1.
\end{aligned}
$$

The $\tilde{q}_i(\cdot, \cdot)$ and $\tilde{r}_i(\cdot, \cdot)$ and $h(\cdot)$ are vector-valued (except for $h(\cdot)$, which is real-valued) Borel functions whose first derivatives with respect to each argument exist. Write $\hat{q}_{i,x}$, $\hat{q}_{i,e}$, $\hat{r}_{i,x}$, $\hat{r}_{i,e}$ for the matrices of first partial derivatives of $\tilde{q}_i(\alpha, \beta)$ and $\tilde{r}_i(\alpha, \beta)$ with respect to the first and second arguments (α and β) evaluated at $\alpha = \hat{x}(t_i)$, $\beta = E\hat{x}(t_i)$, and suppose that they are square integrable. Write $\hat{h}_x$ for the gradient of $h(\alpha)$ evaluated at $\alpha \equiv \hat{x}(T)$, and suppose that it is square integrable.

Define the linear vector-valued operators Q_i, R_i, and scalar-valued operator H, all on the space of square integrable $(n+1)$-dimensional random variables, as follows:

$$Q_i v = E[\hat{q}_{i,x} \cdot v + \hat{q}_{i,e} \cdot Ev],$$

$$R_i v = E[\hat{r}_{i,x} \cdot v + \hat{r}_{i,e} \cdot Ev],$$

$$Hv = E\hat{h}'_x \cdot v,$$

where v is an arbitrary $(n+1)$-vector-valued random variable with square integrable components. Let Q_i^j be the jth component of the vector-valued functional Q_i; i.e., $Q_i^j v = E[(\hat{q}_{i,x}^j)' \cdot v + (E\hat{q}_{i,e}^j)'v]$, where $\hat{q}_{i,x}^j$ is the gradient of $q_i^j(\alpha, \beta)$ evaluated at $\alpha = \hat{x}(t_i)$, $\beta = E\hat{x}(t_i)$, and $Q_i v = \sum_j Q_i^j v$.

For any R^{n+1}-valued random variable v with $E|v|^2 < \infty$, and any generalized sequence $\{y_\alpha\}$ of R^{n+1}-valued square integrable random variables with $\lim_\alpha |y_\alpha - v|^2 = 0$, let

$$\lim_{\substack{y_\alpha \to v \\ \varepsilon \to 0}} \frac{1}{\varepsilon} E[\tilde{q}_i(\hat{x}(t_i) + \varepsilon y_\alpha, E\hat{x}(t_i) + E\varepsilon y_\alpha) - \tilde{q}_i(\hat{x}(t_i), E\hat{x}(t_i))] = Q_i \cdot v$$

as $\varepsilon \to 0$, and similarly for the constraints r_i. Assume that the components of the vector-valued linear functional R_0 are linearly independent, and similarly for those of R_{k+1}.

Assumption 2.5. For the inactive[2] inequality constraints $q_i^j(\cdot)$, suppose that there is some $\varepsilon_i > 0$ so that

$$q_i^j(\hat{x}(t_i) + v) < 0$$

for $E|v|^2 < \varepsilon_i$. For each i suppose that there is a square integrable random variable v_i so that for each active component $q_i^j(\cdot)$ of $q_i(\cdot)$, we have

$$Q_i^j \cdot v_i < 0.$$

3. A variational result of Neustadt. For future reference, we describe a variational result of Neustadt [1]. Let $\mathcal{T}$ denote a locally convex linear topological space, and let Q be a set in $\mathcal{T}$.

DEFINITION. For any integer μ, let P^μ denote the set of vectors in R^μ, $\{\beta : \beta_i \geqq 0, \sum_{i=1}^\mu \beta_i \leqq 1\}$. Let K be a convex set in $\mathcal{T}$ which contains the origin $\{0\}$ and some point other than $\{0\}$. For each μ points $w_1, \cdots, w_\mu$ of K and arbitrary neighborhood N of $\{0\}$, let there exist an $\varepsilon_0 > 0$ (depending on $w_1, \cdots, w_v$ and N) so that, for each ε in $(0, \varepsilon_0]$, there is a continuous map $\zeta_\varepsilon(\beta)$ from P^μ to $\mathcal{T}$ with the property

$$\zeta_\varepsilon(\beta) \subset \left\{ \varepsilon \left(\sum_{i=1}^{\mu} \beta_i w_i + N \right) \right\} \cap Q.$$

Then K is said to be a *first order convex approximation to* Q.

3.1. A basic optimization problem. Let Q' be a set in $\mathcal{T}$. For some finite given integers μ and β, let $\varphi_i(\cdot)$, $-\beta \leqq i \leqq \mu$, be real-valued functions on $\mathcal{T}$. Let C

[2] $q_i^j(\cdot)$ is said to be active if $q_i^j(\hat{x}(t_i)) = 0$. Otherwise $q_i^j(\hat{x}(t_i)) < 0$, and the constraint is said to be inactive.

denote the set of points w in $\mathscr{T}$ which satisfy the constraints $\varphi_i(w) = 0, i = 1, \cdots, \mu$, $\varphi_{-i}(w) \leqq 0$, $i = 1, \cdots, \beta$. A point $\hat{w} \in \mathscr{T}$ is said to be a *local solution* to the optimization problem (or, more loosely, an *optimal solution*) if, for some neighborhood N of $\{0\}$ in $\mathscr{T}$, $\varphi_0(w) \geqq \varphi_0(\hat{w})$ for all w in $(\hat{w} + N) \cap C \cap Q'$. It is desired to find a necessary condition for a point $\hat{w}$ to be an optimal solution. The constraints $\varphi_{-i}(\cdot)$, $i > 0$, for which $\varphi_{-i}(\hat{w}) = 0$ are called the *active constraints*. Define the set of indices $J = \{i: \varphi_{-i}(\hat{w}) = 0, i > 0\} \cup \{0\}$.

The basic necessary condition for optimality. First we collect some assumptions.

Assumption 3.1. The $\varphi_i(\cdot)$, $i \geqq 1$, are continuous at $\hat{w}$. There are continuous and linearly independent functionals $l_1, \cdots, l_\mu$ with the following property. For any element $w \in \mathscr{T}$, and any generalized sequence $\{w_\alpha\}$ converging to w in $\mathscr{T}$, we have

$$\lim_{\substack{\varepsilon \to 0 \\ w_\alpha \to w}} [\varphi_i(\hat{w} + \varepsilon w_\alpha) - \varphi_i(\hat{w})]/\varepsilon = l_i(w) \quad \text{as} \quad \varepsilon \to 0.$$

Assumption 3.2. There is a neighborhood N of $\{0\}$ in $\mathscr{T}$ so that $\varphi_{-i}(\hat{w} + w) < 0$ for $w \in N$, and all *inactive* constraints $\varphi_{-i}(\cdot)$.

Assumption 3.3. Let the active constraints and also $\varphi_0(\cdot)$ be continuous at $\hat{w}$. For the active constraints, let there exist continuous and convex functionals $c_i(\cdot)$ with the property that for any $w \in \mathscr{T}$, and any generalized sequence w_α converging to w in $\mathscr{T}$,

$$\lim_{\substack{w_\alpha \to w \\ \varepsilon \to 0}} [\varphi_{-i}(\hat{w} + \varepsilon w_\alpha) - \varphi_{-i}(\hat{w})]/\varepsilon = c_i(w).$$

Assume that there is some w and some $j \in J$ for which $c_j(w) > 0$. Let there be a w for which $c_j(w) < 0$ for all $j \in J$.

A case of particular importance is that in which the differentials $c_i(\cdot)$ are linear functionals. Then the next to last sentence of Assumption 3.3 is implied by the last sentence of Assumption 3.3.

We now have a particular case of Neustadt [1, Theorem 4.2]. The *local* or *optimal solution* here is called a *totally regular local solution* in [1].

THEOREM 1. *Under Assumptions* 3.1–3.3 *define* $Q \equiv Q' - \hat{w}$. *Let* $\hat{w}$ *be a local solution to the optimization problem. Then there exist* $\alpha_1, \cdots, \alpha_\mu, \alpha_0, \alpha_{-1}, \cdots, \alpha_{-\beta}$, *not all zero, with* $\alpha_{-i} \leqq 0$ *for* $i \geqq 0$ *so that*

$$\sum_{i=1}^{\mu} \alpha_i l_i(w) + \sum_{i \in J} \alpha_{-i} c_i(w) \leqq 0$$

for all w *in* $\bar{K}$, *where* K *is a first order convex approximation to* Q, *and* $\bar{K}$ *is the closure of* K *in* $\mathscr{T}$.

Remark. Let $\varphi_i(\cdot) \equiv 0$, $i > 0$. If there is a $w \in K$ for which $c_j(w) < 0$ for all active j, then $\alpha_0 < 0$, and we can obtain $\alpha_0 = -1$.

3.2. Identification with the stochastic control problem. For the problems of the sequel we define $\mathscr{T}$ to be the locally convex linear topological space of $(n+1)$-dimensional random functions v with values $v(\omega, t)$, where the generalized sequence v_α tends to zero in $\mathscr{T}$ if and only if

$$\lim_{\alpha} E|v_\alpha(\cdot, t)|^2 = 0$$

for each t in $[0, T]$. The set Q' is defined to be the set of solutions[3] $x(\cdot, \cdot)$, $0 \leqq t \leqq T$, to (1) for all controls and initial conditions satisfying Assumptions 2.1, 2.3. The (inequality constraint) functions $\{q_l^j\}$ are identified with the $\{\varphi_{-l}, l > 0\}$ and the (equality constraint) functions $\{r_l^j\}$ with the $\{\varphi_l, l > 0\}$ of § 3. Also $\varphi(x) = Ex_0(T) + Eh(x(T))$. $\hat{x}$ is an optimal element of Q' and $Q \equiv Q' - \hat{x}$. Assumptions 2.1–2.5 imply Assumptions 3.1–3.3.

With the framework of constraints (4), we can include constraints such as $\int_0^{t_i} g_i(x(s))\, ds < d_i$ and can approximate constraints such as $P\{x(t) \in A\} \leqq d_i$, where A has a smooth boundary. More general inequality constraints than (4) can be included, once the appropriate linear or convex differentials c_i (see Assumption 3.3) are calculated.

4. The linearized equations. Consider the Itô stochastic differential equations (6) and (7), where $0 \leqq \tau$ is fixed and t satisfies $\tau \leqq t \leqq T$, and $\Phi(t, \tau)$ is an $(n + 1) \times (n + 1)$ random matrix[4]

$$dy(t) = \hat{f}_x \cdot y(t)\, dt + \sum_i dz_i(t) \hat{\sigma}_{i,x} y(t), \tag{6}$$

$$d\Phi(t, \tau) = \hat{f}_x \Phi(t, \tau)\, dt + \sum_i dz_i(t) \hat{\sigma}_{i,x} \Phi(t, \tau), \tag{7}$$

where, by assumption, $\Phi(\tau, \tau) = I$, the identity, and also by assumption, $E|y(\tau)|^2 < \infty$ and $y(\tau)$ is independent of $z(t) - z(s)$ for all $t \geqq s \geqq \tau$. Both (6) and (7) have unique continuous (in t) solutions with probability 1 (w.p.1), with finite mean square values. We can suppose that the chosen continuous version of $\Phi(t, \tau)$ is measurable in (t, ω) for each τ. The uniqueness of the solution to (6) implies that, for each $\tau \in [0, T]$, w.p.1,

$$\Phi(t, \tau) y(\tau) = y(t)$$

and, for $t > \tau_1 > \tau$, w.p.1,

$$\Phi(t, \tau_1)\Phi(\tau_1, \tau) = \Phi(t, \tau).$$

Furthermore, if we fix t and let τ vary in the range $0 \leqq T \leqq t$, $\Phi(t, \tau)$ is mean square continuous in τ, uniformly in $t \in [\tau, T]$. Indeed, we have w.p.1 that $\Phi(t, \tau + \varepsilon)$ and $\Phi(t, \tau)$ are the solutions (w.p.1) of (7) which start at time $\tau + \varepsilon$ with initial values I and $\Phi(\tau + \varepsilon, \tau)$, respectively. By known estimates for solutions of stochastic differential equations, for some real K_i,

$$E|\Phi(\tau + \varepsilon, \tau) - I|^4 \leqq K_1 \varepsilon^2;$$

and hence for t in $[\tau + \varepsilon, T]$,

$$E|\Phi(t, \tau + \varepsilon) - \Phi(t, \tau)|^4 \leqq K_2 \varepsilon^2. \tag{8}$$

Equation (8) implies that there is a continuous version of $\Phi(T, \tau)$ [14, Proposition

[3] It is easiest to work in the space of random functions $\mathscr{T}$, as it is described above. By Assumptions 2.1 and 2.2 we lose nothing by altering $\mathscr{T}$ so that $v_n \to 0$ if $E|v_n(\omega, t)|^p \to 0$ for any $p \geqq 2$. In this case the quadratic estimates of Assumption 2.4 on q_i and r_i can be replaced by $|q_i(x(t_i))| \leqq K_0(1 + E|x(t_i)|^p)$, etc. More general situations are obviously possible and, in particular, the Lipschitz and growth conditions on the zeroth component of $f(\alpha, \beta, t)$ can be relaxed.

[4] $\hat{f}_x$ denotes the random matrix $f_x(\hat{x}(t), \hat{u}(t), t)$; and similarly for $\hat{\sigma}_{i,x}$.

III.5.3] (τ is the parameter, $0 \leqq \tau \leqq T$). Finally, if $E|y(0)|^2 = O(\varepsilon^2)$ (or $o(\varepsilon^2)$), then $E|y(t)|^2 = O(\varepsilon^2)$ (or $o(\varepsilon^2)$). This last fact will be used frequently in Theorem 2.

5. The convex cone K. We require the following lemma.[5]

LEMMA 1. *Using Assumptions* 2.1 *and* 2.2 *let* $\Phi(\cdot,\cdot)$ *be a* R^{n+1}*-valued measurable function with values* $\phi(\omega, t)$, $0 \leqq t \leqq T$. *Suppose that* $\phi(\omega,\cdot)$ *is Lebesgue integrable on* $[0, T]$ *for almost all fixed* ω. *Then the function* $F(\cdot,\cdot)$ *defined by*

$$F(\omega, t) = \int_{t_0}^{t} \phi(\omega, s)\, ds$$

is differentiable with respect to t *on an* (ω, t) *set of full measure with derivative* $\phi(\omega, t)$. *Thus, there is a null set* $T_1 \subset (0, T)$ *so that, at each* $t \notin T_1$, $F(\omega, t)$ *is differentiable with derivative* $\phi(\omega, t)$, *w.p.*1. *In particular, if we define* $\phi(\cdot,\cdot)$ *by* $\phi(\omega, s) = f(\hat{x}(s), \hat{u}(s), s)$ *and let* α_1, α_2 *be any scalars, we have*

$$\frac{1}{\varepsilon}\int_{t-\varepsilon\alpha_1}^{t+\varepsilon\alpha_2} f(\hat{x}(s), \hat{u}(s), s)\, ds - (\alpha_2 + \alpha_1) f(\hat{x}(t), \hat{u}(t), t) \to 0$$

*w.p.*1, *for any* t *not in some null set* T_1.

There is a null set $T_2 \subset (0, T)$ *so that for any* $t \notin T_2$ *and any* R^{n_1}*-valued random variable* v,

$$\frac{1}{\varepsilon}\int_{t-\varepsilon\alpha_1}^{t+\varepsilon\alpha_2} f(\hat{x}(s), v, s)\, ds - (\alpha_2 + \alpha_1) f(\hat{x}(t), v, t) \to 0$$

*w.p.*1.

Proof. For arbitrary scalars α_1, α_2, define the function $F_r(\cdot,\cdot)$:

$$F_r(\omega, t) = \frac{1}{r}\int_{t-\alpha_1 r}^{t+\alpha_2 r} \phi(\omega, s)\, ds,$$

where r is rational in $[0, 1]$. There is a null ω-set N_0 so that, for $\omega \notin N_0$, $\phi(\omega,\cdot)$ is Lebesgue integrable on $[0, T]$ and, hence, for $\omega \notin N_0$,

$$F_r(\omega, t) - (\alpha_1 + \alpha_2)\phi(\omega, t) \to 0$$

for almost all t (the null t-set depending on ω). Also $F_r(\omega, t)$ converges to $(\alpha_1 + \alpha_2)\phi(\omega, t)$ on a measurable set $S \subset (\Omega - N_0) \times [0, T]$ as $r \to 0$. If $F_r(\omega, t)$ converges as $r \to 0$ through the rational numbers, it converges to the same limit as $r \to 0$ through any sequence.

The Lebesgue measure of the fixed ω-sections of S (for $\omega \notin N_0$) is T. Hence by Fubini's theorem, the measurable set S has full measure. Thus, there is a null t-set T_1 so that for $t \notin T_1$, $F_r(\omega, t) \to (\alpha_1 + \alpha_0)\phi(\omega, t)$ w.p.1. The statements of the first paragraph of the lemma follow from this.

Let $g(\cdot,\cdot)$ denote a Borel function from $R^m \times [0, T]$ to R^{n+1}, where $g(\cdot, t)$ is continuous, uniformly in t. Let $g(\gamma(\cdot),\cdot)$ be integrable on $[0, T]$ for any R^m-valued continuous function $\gamma(\cdot)$ on $[0, T]$. Then there is a null set T_2 so that, for $t \notin T_2$ and any continuous R^m-valued function $\gamma(\cdot)$,

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\int_{t-\varepsilon\alpha_1}^{t+\varepsilon\alpha_2} g(\gamma(s), s)\, ds - (\alpha_1 + \alpha_2) g(\gamma(t), t) = 0.$$

[5] The proof of Lemma 1 resulted from a discussion with W. Fleming.

Let $\rho = (\rho_1, \rho_2)$, where ρ_1 is $(n+1)$-dimensional, and ρ_2 is n_1-dimensional. Define $g(\cdot, \cdot)$ by $g(\rho, t) = f(\rho_1, \rho_2, t)$. The second paragraph of the lemma then follows by noting that, for almost all ω, $\hat{x}(\omega, \cdot)$ is continuous in t and $g((\hat{x}(\omega, t), v(\omega)), t) = f(\hat{x}(\omega, t), v(\omega), t)$.

The convex cone K. For any fixed s in the set $[0, T] - \hat{T}$, where[6] $\hat{T} = T_1 \cup T_2$, and any random variable u_s which is $\mathscr{B}_s$ measurable and has values in $\mathscr{U}_s$, define the element $\delta x_{s,u_s}$ of $\mathscr{T}$. Following our usual notation, we use $\delta x_{s,u_s}(t)$ for either $\delta x_{s,u_s}(\cdot, t)$ or $\delta x_{s,u_s}(\omega, t)$:

$$\delta x_{s,u_s}(t) = \begin{cases} 0, & 0 \leqq t < s \leqq T, \\ \Phi(t, s)[f(\hat{x}(s), u_s, s) - f(\hat{x}(s), \hat{u}(s), s)], & T \geqq t \geqq s. \end{cases}$$

Let K denote the set of convex finite combinations of functions of the type $c_0\Phi(\cdot, 0)\delta x(0)$, where c_0 is arbitrary in $[0, \infty)$ and $\delta x(0)$ is an arbitrary but admissible initial condition (i.e., $\delta x(0)$ is $\mathscr{B}_0$ measurable, $\mathscr{U}_0$-valued and $E|\delta x(0)|^2 < \infty$), and functions of the type $c\delta x_{s,u_s}$, where c is arbitrary in $[0, \infty)$ and s is arbitrary in $[0, T] - \hat{T}$, and u_s is an arbitrary $\mathscr{B}_s$-measurable and $\mathscr{U}_s$-valued random variable. Define

$$\delta x_0(t) = \Phi(t, 0)\delta x(0).$$

By Theorem 2, K is a first order convex approximation to the set $Q' - \hat{x} \equiv Q$.

THEOREM 2. *Under Assumptions* 2.1–2.3, *K is a first order convex approximation to* $Q \equiv Q' - \hat{x}$.

Proof. Let m denote an arbitrary, but fixed, integer. Define the set

$$\Lambda \equiv \{\lambda = (\lambda_1, \cdots, \lambda_m) : \lambda_i \geqq 0, \sum_i \lambda_i \leqq 1\}.$$

Let $\delta x^1, \cdots, \delta x^m$ denote any m elements of K. Then, there is an integer q, a set of fixed times s_i, $i = 1, \cdots, q$, a set of $\mathscr{U}_{s_i}$-valued and $\mathscr{B}_{s_i}$-measurable random variables u_{s_i} (written u_i), $i = 1, \cdots, q$, and a set of $\tilde{\beta}_{ij} \geqq 0$, $\beta_{ij} \geqq 0$, where $\sum_j (\beta_{ij} + \tilde{\beta}_{ij}) \leqq 1$, and admissible initial conditions $\delta x_0^i(0)$, $i = 1, \cdots, q$, so that each δx^i has the representation

$$\delta x^i = \sum_{j=1}^{q} \beta_{ij}\, \delta x_{s_j,u_j} + \sum_{j=1}^{q} \tilde{\beta}_{ij}\, \delta x_0^j.$$

We assume that $s_i \leqq s_{i+1}$. Any element in $\tilde{K}$, the convex hull of $(0, \delta x^1, \cdots, \delta x^m)$, corresponds to some $\lambda \in \Lambda$ (and conversely) and has the form

$$\begin{aligned} \delta x_\lambda &= \sum_{i=1}^{m} \lambda_i \left(\sum_{j=1}^{q} \beta_{ij}\, \delta x_{s_j,u_j} \right) + \sum_{i=1}^{m} \lambda_i \left(\sum_{j=1}^{q} \tilde{\beta}_{ij}\, \delta x_0^j \right) \\ &= \sum_{j=1}^{q} \delta t_j(\lambda) \delta x_{s_j,u_j} + \sum_{j=1}^{q} \delta \tilde{t}_j(\lambda) \delta x_0^j, \\ \delta t_j(\lambda) &= \sum_{i=1}^{m} \beta_{ij}\lambda_i, \qquad \delta \tilde{t}_j(\lambda) = \sum_{i=1}^{m} \tilde{\beta}_{ij}\lambda_i. \end{aligned}$$

Note that $\varepsilon \delta t_i(\lambda) = \delta t_i(\varepsilon\lambda)$ for any scalar ε in $(0, \sum_i \lambda_i]$, and similarly for $\delta \tilde{t}_i(\lambda)$.

[6] The T_i are defined in Lemma 1.

Let $\tilde{O}(\varepsilon)$ denote any random function v_ε for which $E^{1/2}|v_\varepsilon(t)|^2 = O(\varepsilon)$ for each $t \in [0, T]$, and write $v_\varepsilon = \tilde{o}(\varepsilon)$ if $E^{1/2}|v_\varepsilon(t)|^2 = o(\varepsilon)$ for each $t \in [0, T]$. To prove the theorem we must show that there is an $\varepsilon_0 > 0$ so that, for each $\varepsilon < \varepsilon_0$, there is a continuous map $\zeta_\varepsilon(\lambda)$ from Λ into Q of the form

$$\zeta_\varepsilon(\lambda) = \varepsilon\, \delta x_\lambda + \rho_{\varepsilon,\lambda}, \tag{9}$$

where $\rho_{\varepsilon,\lambda} = \tilde{o}(\varepsilon)$; of course, there is a corresponding map $\zeta_\varepsilon(\lambda) + \hat{x}$ from Λ into Q'.

Next, a perturbed control and perturbed initial condition will be described. Suppose first that the s_i are distinct and $s_i \notin \hat{T} = T_1 \cup T_2$. Define

$$\tau \equiv \sup_{i,\lambda\in\Lambda} \delta t_i(\lambda) \cdot q.$$

There is an $\varepsilon_1 > 0$ so that for $\varepsilon < \varepsilon_1$, all $s_i - \varepsilon\tau \geqq 0$ and $\varepsilon\tau \leqq \min[\delta(s_1), \cdots, \delta(s_q)]$. For $\varepsilon < \varepsilon_1$ and all $\lambda \in \Lambda$, define the sets

$$I_i(\varepsilon\lambda) \equiv \{t : s_i - \varepsilon\delta t_i(\lambda) < t \leqq s_i\}, \qquad i = 1, \cdots, q.$$

There is an ε_0, $\varepsilon_1 \geqq \varepsilon_0 > 0$, so that the $I_1(\varepsilon\lambda), \cdots, I_q(\varepsilon\lambda)$ are pairwise disjoint for $\varepsilon < \varepsilon_0$ and any $\lambda \in \Lambda$. Define the perturbed control

$$u_{\varepsilon\lambda}(t) = \begin{cases} \hat{u}(t), & t \notin \bigcup_i I_i(\varepsilon\lambda), \\ \tilde{u}_{s_i - \varepsilon\tau}, & t \in I_i(\varepsilon\lambda), \end{cases} \tag{10}$$

where $\tilde{u}_{s_i-\varepsilon\tau}$ corresponds to u_{s_i} by Assumption 2.3, and as $\varepsilon \to 0$, (5) of Assumption 2.3 holds.

If the s_i are not distinct, we follow the method for the deterministic problem [16] and define τ_i by

$$\tau_i = \begin{cases} \delta t_i(\lambda) + \cdots + \delta t_q(\lambda) & \text{if } s_i = s_{i+1} = \cdots = s_q, \\ \delta t_i(\lambda) + \cdots + \delta t_r(\lambda) & \text{if } s_i = s_{i+1} = \cdots = s_r < s_{r+1}, r < q, \end{cases}$$

and $I_i(\varepsilon\tau)$ by

$$\begin{aligned} I_i(\varepsilon\lambda) &= \{t : s_i - \varepsilon\tau_i < t \leqq s_i - \varepsilon\tau_i + \varepsilon\, \delta t_i(\lambda)\} \\ &= \{t : s_i - \varepsilon(\delta t_i(\lambda) + \cdots + \delta t_r(\lambda)) < t \leqq s_i - \varepsilon(\delta t_{i+1}(\lambda) + \cdots + \delta t_r(\lambda))\}. \end{aligned}$$

Then define $u_{\varepsilon\lambda}(t)$ as in (10). Thus, if some s_i are identical, the intervals are shifted to the left.

By Assumptions 2.1 and 2.3 the perturbed control $u_{\varepsilon\lambda}$ is admissible. Let $x_{\varepsilon\lambda} \in \mathscr{T}$ denote the solution of (1) for control $u_{\varepsilon\lambda}$ and initial condition

$$\hat{x}(0) + \varepsilon \sum_{j=1}^{q} \delta\tilde{t}_j(\lambda)\, \delta x_0^j(0) \equiv \hat{x}(0) + \varepsilon\, \delta x_\lambda(0) \equiv x_{\varepsilon\lambda}(0), \tag{11}$$

where we use $\varepsilon\, \delta x_\lambda(0) = \delta x_{\varepsilon\lambda}(0)$. Define

$$\zeta_\varepsilon(\lambda) = x_{\varepsilon\lambda}. \tag{12}$$

Fix ε in $(0, \varepsilon_0)$. Let $\lambda(n) \to \lambda$ in Λ as $n \to \infty$. Then $E|x_{\varepsilon\lambda(n)}(0) - x_{\varepsilon\lambda}(0)|^2 \to 0$, and the total length of the intervals on which $u_{\varepsilon\lambda(n)}(t) \neq u_{\varepsilon\lambda}(t)$ converges to zero.

These facts imply that $E|x_{\varepsilon\lambda(n)}(t) - x_{\varepsilon\lambda}(t)|^2 \to 0$ for each t, which implies the continuity of $\zeta_\varepsilon(\lambda)$ for each $\varepsilon < \varepsilon_0$. We need only prove the expansion (9), and this will be done in three parts.

(a) Let the K_i be real numbers, and define the process $y_{\varepsilon\lambda}(\cdot)$ as the solution to (13c). Equations (13a), (13b) hold.

$$d\hat{x}(t) = f(\hat{x}(t), \hat{u}(t), t)\,dt + \sum_j dz_j(t)\sigma_j(\hat{x}(t), t), \tag{13a}$$

$$dx_{\varepsilon\lambda}(t) = f(x_{\varepsilon\lambda}(t), u_{\varepsilon\lambda}(t), t)\,dt + \sum_j dz_j(t)\sigma_j(x_{\varepsilon\lambda}(t), t), \tag{13b}$$

$$\begin{aligned} dy_{\varepsilon\lambda}(t) =& \hat{f}_x y_{\varepsilon\lambda}(t)\,dt + [f(\hat{x}(t), u_{\varepsilon\lambda}(t), t) - f(\hat{x}(t), \hat{u}(t), t)]\,dt \\ &+ \sum_j dz_j(t)\hat{\sigma}_{j,x} y_{\varepsilon\lambda}(t), \\ y_{\varepsilon\lambda}(0) &= \delta x_{\varepsilon\lambda}(0) = \varepsilon\,\delta x_\lambda(0). \end{aligned} \tag{13c}$$

Using standard estimates for solutions of Itô stochastic differential equations it can be shown that, for some $K_1 < \infty$,

$$\max_{\varepsilon<\varepsilon_0, \lambda\in\Lambda} E \max_{0\le t\le T} |x_{\varepsilon\lambda}(t)|^2 \le K_1. \tag{14}$$

Next, we define $\tilde{x}(t) \equiv \hat{x}(t) - x_{\varepsilon\lambda}(t)$ and show that

$$E|\tilde{x}(t)|^2 \equiv E|\hat{x}(t) - x_{\varepsilon\lambda}(t)|^2 = O(\varepsilon^2) \tag{15}$$

uniformly in t. Equation (15) holds for $t = 0$. Assume it holds for $t = t_0$ and that $u_{\varepsilon\lambda}(t) = \hat{u}(t)$ for $t \in [t_0, t_1]$. We shall show that (15) holds uniformly in $[t_0, t_1]$. Then, if (15) holds at $t = s_i - \varepsilon\rho$, we show that it holds uniformly in $[s_i - \varepsilon\rho, s_i]$ for any real ρ for which $s_i - \varepsilon\rho \ge 0$. These two facts imply (15) as asserted. Let $\tilde{x}(t) \equiv x_{\varepsilon\lambda}(t) - \hat{x}(t)$. Then,

$$\begin{aligned} \tilde{x}(t) = \tilde{x}(t_0) &+ \int_{t_0}^t [f(x_{\varepsilon\lambda}(s), \hat{u}(s), s) - f(\hat{x}(s), \hat{u}(s), s)]\,ds \\ &+ \int_{t_0}^t [\sigma(x_{\varepsilon\lambda}(s), s) - \sigma(\hat{x}(s), s)]\,dz(s), \end{aligned}$$

where $E|\tilde{x}(t_0)|^2 = O(\varepsilon^2)$. By standard estimates for solutions of stochastic differential equations, we have

$$E|\tilde{x}(t)|^2 \le K_2|\tilde{x}(0)|^2 + K_2 \int_{t_0}^t E|\tilde{x}(s)|^2\,ds,$$

which implies (15) in $[t_0, t_1]$. Next, write

$$\begin{aligned} \tilde{x}(t) = \tilde{x}(s_i - \varepsilon\rho) &+ \int_{s_i-\varepsilon\rho}^t [f(x_{\varepsilon\lambda}(s), u_{\varepsilon\lambda}(s), s) - f(\hat{x}(s), \hat{u}(s), s)]\,ds \\ &+ \int_{s_i-\varepsilon\rho}^t [\sigma(x_{\varepsilon\lambda}(s), s) - \sigma(\hat{x}(s), s)]\,dz(s). \end{aligned}$$

Using the Lipschitz condition on σ, and Schwarz's inequality on the drift term

gives, for $t \geqq s_i - \varepsilon\rho$,

$$E|\tilde{x}(t)|^2 \leqq K_3 E|\tilde{x}(s_i - \varepsilon\rho)|^2 + K_3(t - (s_i - \varepsilon\rho)) \int_{s_i - \varepsilon\rho}^{t} E[f(x_{\varepsilon\lambda}(s), u_{\varepsilon\lambda}(s), s) - f(\hat{x}(s), \hat{u}(s), s)]^2 \, ds + K_3 \int_{s_i - \varepsilon\rho}^{t} |\tilde{x}(s)|^2 \, ds.$$

Using (14) and the growth condition $|f|^2 \leqq K_0(1 + |x|^2)$ in Assumption 2.2 we have

$$E|\tilde{x}(t)|^2 \leqq K_3 E|\tilde{x}(s_i - \varepsilon\rho)|^2 + K_4(t - (s_i - \varepsilon\rho))^2 + K_3 \int_{s_i - \varepsilon\rho}^{t} E|\tilde{x}(s)|^2 \, ds$$

from which (15) follows in $[s_i - \varepsilon\rho, s_i]$.

By reasoning close to the foregoing, it can be shown that

$$E|y_{\varepsilon\lambda}(t)|^2 = O(\varepsilon^2) \tag{16}$$

uniformly in $t \in [0, T]$.

(b) Next, it will be shown that

$$E|x_{\varepsilon\lambda}(t) - \hat{x}(t) - y_{\varepsilon\lambda}(t)|^2 = o(\varepsilon^2) \tag{17}$$

by the method used to show (15). Suppose $\hat{u}(t) = u_{\varepsilon\lambda}(t)$ in $t \in [t_0, t_1]$ and (17) holds for $t = t_0$. Write $\tilde{y}(t) \equiv x_{\varepsilon\lambda}(t) - \hat{x}(t) - y_{\varepsilon\lambda}(t)$. Then, for $t \in [t_0, t_1]$,

$$\begin{aligned} \tilde{y}(t) &= \tilde{y}(t_0) + \int_{t_0}^{t} [f(x_{\varepsilon\lambda}(s), \hat{u}(s), s) - f(\hat{x}(s), \hat{u}(s), s) - \hat{f}_x y(s)] \, ds \\ &\quad + \int_{t_0}^{t} \sum_j dz_j(s)[\sigma_j(x_{\varepsilon\lambda}(s), s) - \sigma_j(\hat{x}(s), s) - \hat{\sigma}_{j,x} y_{\varepsilon\lambda}(s)] \qquad (18) \\ &= \tilde{y}(t_0) + \int_{t_0}^{t} \hat{f}_x \tilde{y}(s) \, ds + \int_{t_0}^{t} \sum_j dz_j(s) \hat{\sigma}_{j,x} \tilde{y}(s) + e_1(t) + e_2(t), \end{aligned}$$

where, for $\tilde{x}(s) \equiv x(s) - \hat{x}(s)$, we define

$$e_1(t) = \int_{t_0}^{t} [f_x(\hat{x}(s) + \varphi(s)\tilde{x}(s), \hat{u}(s), s) - f_x(\hat{x}(s), \hat{u}(s), s)]\tilde{x}(s) \, ds,$$

$$e_2(t) = \int_{t_0}^{t} \sum dz_j(s)[\sigma_{j,x}(\hat{x}(s) + \tilde{\varphi}(s)\tilde{x}(s), s) - \sigma_{j,x}(\hat{x}(s), s)]\tilde{x}(s),$$

where $\varphi(\cdot)$ and $\tilde{\varphi}(\cdot)$ are scalar-valued random functions with values in $[0, 1]$. By (15) and the continuity (in α) and boundedness properties of $f_x(\alpha, \beta, s)$ and $\sigma_{i,x}(\alpha, s)$,

$$E|e_i(t)|^2 = o(\varepsilon^2)$$

uniformly in t. With this estimate, (17) easily follows from (18) in $[t_0, t_1]$.

Next write $\delta t_i(\lambda) = \rho_i$ and let $E|\tilde{y}(s - \varepsilon\tau_i)|^2 = o(\varepsilon^2)$. For $t \in [s_i - \varepsilon\tau_i, s_i - \varepsilon\tau_i + \varepsilon\rho_i]$ write

$$\begin{aligned}\tilde{y}(t) = \tilde{y}(s_i - \varepsilon\tau_i) &+ \int_{s_i-\varepsilon\tau_i}^{t} [f(x_{\varepsilon\lambda}(s), u_{\varepsilon\lambda}(s), s) - f(\hat{x}(s), \hat{u}(s), s) - \hat{f}_x y_{\varepsilon\lambda}(s)]\, ds \\ &- \int_{s_i-\varepsilon\tau_i}^{t} [f(\hat{x}(s), u_{\varepsilon\lambda}(s), s) - f(\hat{x}(s), \hat{u}(s), s)]\, ds \\ &+ \int_{s_i-\varepsilon\tau_i}^{t} \sum_j dz_j(s)[\sigma_i(x_{\varepsilon\lambda}(s), s) - \sigma_i(\hat{x}(s), s) - \hat{\sigma}_{i,x} y_{\varepsilon\lambda}(s)].\end{aligned} \tag{19}$$

Equation (19) can be written as

$$\begin{aligned}\tilde{y}(t) = \tilde{y}(s_i - \varepsilon\tau_i) &+ \int_{si-\varepsilon\tau_i}^{t} [f(x_{\varepsilon\lambda}(s), u_{\varepsilon\lambda}(s), s) - f(\hat{x}(s), u_{\varepsilon\lambda}(s), s)] \\ &+ \int_{s_i-\varepsilon\tau_i}^{t} \sum_j dz_j(s)[\sigma_j(x_{\varepsilon\lambda}(s), s) - \sigma_j(\hat{x}(s), s)] + e_3(t),\end{aligned} \tag{20}$$

where we define

$$e_3(t) = -\left[\int_{s_i-\varepsilon\tau_i}^{t} \hat{f}_x y_{\varepsilon\lambda}(s)\, ds + \int_{s_i-\varepsilon\tau_i}^{t} \sum_j dz_j(s)\hat{\sigma}_{j,x} y_{\varepsilon\lambda}(s)\right].$$

Using $E|y_{\varepsilon\lambda}(s)|^2 = O(\varepsilon^2)$ uniformly in s we get, for t in the desired interval,

$$E\left|\int_{s_i-\varepsilon\tau_i}^{t} \sum_j dz_j(s)\hat{\sigma}_{j,x} y_{\varepsilon\lambda}(s)\right|^2 \leqq K_5 \int_{s_i-\varepsilon\tau_i}^{t} |y_{\varepsilon\lambda}(s)|^2\, ds = o(\varepsilon^2),$$

and similarly for the first term of $e_3(t)$. Using this and the estimates for the two integrals in (20),

$$\begin{aligned}E\left|\int_{s_i-\varepsilon\tau_i}^{t} \sum_j dz_j(s)[\sigma_j(x_{\varepsilon\lambda}(s), s) - \sigma_j(\hat{x}(s), s)]\right|^2 &\\ + E\left|\int_{s_i-\varepsilon\tau_i}^{t} [f(x_{\varepsilon\lambda}(s), u_{\varepsilon\lambda}(s), s) - f(\hat{x}(s), u_{\varepsilon\lambda}(s), s)]\, ds\right|^2 &\\ \leqq K_5 \int_{s_i-\varepsilon\tau_i}^{t} E|\tilde{x}(s)|^2\, ds &\end{aligned}$$

and (15), we have (17) in $[s_i - \varepsilon\tau_i, s_i - \varepsilon\tau_i + \varepsilon\rho_i]$.

(c) Suppose that (21) holds for $t \in [0, T]$:

$$E|y_{\varepsilon\lambda}(t) - \delta x_{\varepsilon\lambda}(t)|^2 = o(\varepsilon^2). \tag{21}$$

Then (17) and (21) imply that $E|x_{\varepsilon\lambda}(t) - \hat{x}(t) - \delta x_{\varepsilon\lambda}(t)|^2 = o(\varepsilon^2)$ uniformly in t in $[0, T]$. Then to complete the proof it is only necessary to show that $x_{\varepsilon\lambda} \in Q'$. But $u_{\varepsilon\lambda}(\cdot)$ satisfies the conditions for admissibility in Assumptions 2.1 and 2.3, $x_{\varepsilon\lambda}(0)$ satisfies the required conditions on $x(0)$ in Assumptions 2.1 and then $x_{\varepsilon\lambda} \in Q'$ by (13b). Thus, we only need to prove (21).

Equation (21) holds for $t = 0$, and indeed, (21) is zero for $t \in [0, s_1 - \varepsilon\tau_1]$. If (21) holds at t_0, then it is true in $[t_0, t_1]$ if $u_{\varepsilon\lambda}(t) = \hat{u}(t)$ in $[t_0, t_1]$, since, w.p.1,

$$y_{\varepsilon\lambda}(t_1) - \delta x_{\varepsilon\lambda}(t_1) = \Phi(t_1, t_0)[y_{\varepsilon\lambda}(t_0) - \delta x_{\varepsilon\lambda}(t_0)]. \tag{22}$$

Next, for $t \in [s_i - \varepsilon\tau_i, s_i - \varepsilon\tau_i + \varepsilon\rho_i]$, where $\rho_i = \delta t_i(\lambda)$,

$$y_{\varepsilon\lambda}(t) = y_{\varepsilon\lambda}(s_i - \varepsilon\tau_i) + J_i^\varepsilon(t) + \tilde{o}(\varepsilon), \tag{23}$$

where we define

$$J_i^\varepsilon(t) = \int_{s_i - \varepsilon\tau_i}^{t} [f(\hat{x}(s), u_{\varepsilon\lambda}(s), s) - f(\hat{x}(s), \hat{u}(s), s)]\, ds.$$

Let $J_i^\varepsilon \equiv J_i^\varepsilon(s_i - \varepsilon\tau_i + \varepsilon\rho_i)$. If $y_{\varepsilon\lambda}^1(t_0) = y_{\varepsilon\lambda}(t_0) + \tilde{o}(\varepsilon)$, and $u_{\varepsilon\lambda}(t) = \hat{u}(t)$, $t \in [t_0, t_1]$, then, w.p.1.,

$$\tilde{y}_{\varepsilon\lambda}^1(t_1) = \Phi(t_1, t_0) y_{\varepsilon\lambda}(t_0) + \tilde{o}(\varepsilon).$$

Furthermore, $E|J_i^\varepsilon(t)|^2 = O(\varepsilon^2)$ uniformly in t, and $\Phi(t - \varepsilon\varphi_1, \tau - \varepsilon\varphi_2) \to \Phi(t, \tau)$ in probability as $\varepsilon \to 0$ for any constants φ_1, φ_2.

The last paragraph implies that w.p.l., for $t \notin \cup I_i^\circ(\varepsilon\lambda)$, where $I_i^\circ(\varepsilon\lambda)$ is the interior of $I_i(\varepsilon\lambda)$, that

$$y_{\varepsilon\lambda}(t) = \sum_{t > s_i} \Phi(t, s_i) J_i^\varepsilon + \Phi(t, 0)\delta x_{\varepsilon\lambda}(0) + \tilde{o}(\varepsilon). \tag{24}$$

Define

$$J_i = \int_{s_i - \varepsilon\tau_i}^{s_i - \varepsilon\tau_i + \varepsilon\rho_i} [f(\hat{x}(s), u_i, s) - f(\hat{x}(s), \hat{u}(s), s)]\, ds.$$

Then Assumption 2.3 implies that $E|J_i^\varepsilon - J_i|^2 = o(\varepsilon^2)$. Thus (24) is valid for J_i replacing J_i^ε. By Lemma 1 (letting $\alpha_2 = -\tau_i + \rho_i, \alpha_1 = \tau_i$),

$$\frac{1}{\varepsilon\rho_i} J_i \to f(\hat{x}(s_i), u_i, s_i) - f(\hat{x}(s_i), \hat{u}(s_i), s_i)$$

w.p.1. as $\varepsilon \to 0$.

Thus, for $t \notin \cup I_i^\circ(\varepsilon\lambda)$,

$$\begin{aligned} y_{\varepsilon\lambda}(t) &= \sum_{t > s_i} \Phi(t, s_i)\delta t_i(\lambda)[f(\hat{x}(s_i), u_i, s_i) - f(\hat{x}(s_i), \hat{u}(s_i), s_i)] \\ &\quad + \Phi(t, 0)\delta x_{\varepsilon\lambda}(0) + \tilde{o}(\varepsilon) \\ &= \delta x_{\varepsilon\lambda}(t) + \tilde{o}(\varepsilon). \end{aligned} \tag{25}$$

Since the sets $I_i^\circ(\varepsilon\lambda)$ decrease to the empty set as $\varepsilon \to 0$, (25) holds for all $t \in [0, T]$.

6. The maximum principle. Combining Theorems 1 and 2 we obtain Theorem 3. Define the $(n+1)$-dimensional column vector $P \equiv (1, 0, \cdots, 0)'$. Theorem 3 reduces to the Pontryagin maximum principle, if the noise is absent ($\sigma \equiv 0$).

THEOREM 3. *Under Assumptions* 2.1–2.5 *there are continuous (in* t*) versions of* $\Phi(T, \cdot), \Phi(t_i, \cdot)$ *(for* $t \leqq T$ *and* $t \leqq t_i$*, respectively). There is a scalar* $\theta \geqq 0$*, vectors* $\alpha_i \leqq 0$, $i = 0, 1, \cdots, k+1$ *(nonpositive components* α_i^j*), where* $\alpha_i^j = 0$ *if* $q_i^j(\hat{x}) < 0$*, and vectors* β_0, β_{k+1} *not all zero, and a null set* $\tilde{T} \in [0, T]$ *so that for all* $t \notin \tilde{T}$*, and all* $\mathscr{B}_t$*-measurable,* $\mathscr{U}_t$*-valued random variables* u_t*, and admissible* $x(0)$, (26a) and (26b) *hold, w.p.*1:

$$\begin{aligned} &\Big\{\theta E[P + h_x(\hat{x}(T))]'\Phi(T, t) + \sum_{i: t_i > t} \alpha_i' E[\hat{q}_{i,x} + E\hat{q}_{i,e}]\Phi(t_i, t) \\ &\quad + \beta_{k+1}' E[\hat{r}_{k+1,x} + E\hat{r}_{k+1,e}]\Phi(T, t)\Big\} \cdot \{f(\hat{x}(t), u_t, t) - f(\hat{x}(t), \hat{u}(t), t)\} \leqq 0, \end{aligned} \tag{26a}$$

$$\Big\{\theta E[P + h_x(\hat{x}_T)]'\Phi(T,0) + \sum_i \alpha_i' E[\hat{q}_{i,x} + E\hat{q}_{i,e}]\Phi(t_i,0)$$

(26b)

$$+ \beta_{k+1}' E[\hat{r}_{k+1,x} + E\hat{r}_{k+1,e}]\Phi(T,0) + \beta_0' E[\hat{r}_{0,x} + E\hat{r}_{0,e}]\Big\}\delta x(0) \leqq 0.$$

Inequality (26b) *implies the term in braces in* (26b) *is zero. Define the vector* $p(T)$ *by its transpose*:

(27) $$p'(T) = \theta[P + h_x(\hat{x}(T))]' + \beta_{k+1}'[\hat{r}_{k+1,x} + E\hat{r}_{k+1,e}] + \alpha_{k+1}'[\hat{q}_{T,x} + E\hat{q}_{T,e}].$$

Define the $(n+1)$*-dimensional random function* $p(t)$, $t < T$, *by its transpose*:

$$p'(t) = p'(T)\Phi(T,t), \qquad t_k \leqq t \leqq t_{k+1} = T,$$

(28) $$p'(t_i^-) = p'(t_i) + \alpha_i'[\hat{q}_{i,x} + E\hat{q}_{i,e}], \qquad i = 1, \cdots, k,$$

$$p'(t) = p'(t_i^-)\Phi(t_i,t), \qquad 0 = t_0 \leqq t_{i-1} \leqq t < t_i.$$

With the use of (27) *and* (28), (26) *can be written as*

(29a) $$Ep'(t)[f(\hat{x}(t), u_t, t) - f(\hat{x}(t), \hat{u}(t), t)] \leqq 0,$$

(29b) $$E[p'(0) + \beta_0'(\hat{r}_{0,x} + E\hat{r}_{0,e})]\delta x(0) = 0.$$

*Furthermore, w.p.*1,

(30a) $$E\{p'(t)[f(x(t), u(t), t) - f(x(t), u_t, t)]|\mathscr{B}_t\} \leqq 0,$$

(30b) $$E\{[p'(0) + \alpha_0'(\hat{r}_{0,x} + E\hat{r}_{0,e})]|\mathscr{B}_0\} = 0.$$

Proof. The proof of (26) follows from Theorem 1 using the appropriate identification of the (continuous in $\mathscr{T}$ by Assumption 2.4) components of Q_i, R_i with the c_j, l_j in Theorem 1, and the fact that K is a first order convex approximation to $Q' = Q - \hat{x}$ by Theorem 2. Also the (continuous in $\mathscr{T}$ by Assumption 2.4) linear operator which acts on $\delta x(T)$ in $E[P + h_x(\hat{x}(T))]'\delta x(T)$ is identified with c_0. Equations (29) follow from (26) upon using the substitution (27), (28). To prove (30a) suppose that (30a) is violated on a $\mathscr{B}_t$-measurable set B_t with $P(B_t) > 0$. Define $\bar{u}_t = u_t$ on B_t, $\bar{u}_t = \hat{u}(t)$ on $\Omega - B_t$. Then (29a) is violated with the admissible $\bar{u}_t$ replacing the u_t there. A similar proof yields (30b).

7. Extensions to closed loop systems. Thus far the admissible controls are defined to be measurable on the a priori given σ-algebra $\mathscr{B}_t$. If the admissible controls are assumed to depend explicitly on the state—or on its past values, i.e., $u(t) = u(x(t), t)$ or $u(t) = u(x_s, s \leqq t, t)$—then a very similar development can be carried out provided either the Lipschitz condition,

(31) $$|u(\alpha, t) - u(\beta, t)| \leqq K|\alpha - \beta|,$$

or the generalized Lipschitz condition,

$$|u(x,t) - u(y,t)| \leqq \int_0^t |x(t-s) - y(t-s)|\, dm(s),$$

for a bounded measure $m(\cdot)$ holds.[7] Indeed, with the use of the perturbed controls

[7] For more detail on the more general stochastic differential delay system, see [15].

and a convex cone K of the type used in Theorem 2, we obtain Theorem 3, with the exception that the $\hat{f}_x$ terms in the $y_{\varepsilon\lambda}(t)$ and $\Phi(t, \tau)$ equations are replaced by $\hat{f}_x + \hat{f}_u \cdot \hat{u}_x$. In particular, let the data available to the control at time t be $g(x(t), t)$, where $g(\cdot, \cdot)$ is a Borel function satisfying (31) with values in some Euclidean space and $|g(\alpha, t)|^2 \leqq K_0(1 + |\alpha|^2)$. Let the class of admissible controls $\tilde{\mathscr{U}}$ be the family of Borel functions $u(\cdot, \cdot)$ with values $u(g(x(t), t), t)$ and which satisfies (31), and which has values in $\mathscr{U}_t$ at time t. Let $x(0)$ satisfy the relevent parts of Assumption 2.1. Let Assumption 2.2 hold for admissible u. The convex cone is composed of elements with values (for almost all s)

$$\Phi(t, 0)\delta x(0) + \Phi(t, s)[f(\hat{x}(s), u(g(\hat{x}(s), s), s), s) - f(\hat{x}(s), \hat{u}(g(\hat{x}(s), s), s), s)].$$

For each t and admissible u, it is supposed that there is a continuous function $\tilde{u}(\cdot, \cdot)$ (of g, t) satisfying (31) with $\tilde{u}(g, t) \in \mathscr{U}_t$ for all g and $|\tilde{u}(g, t)|^2 \leqq K_0(1 + |g|^2)$ such that $u(g, s) = \tilde{u}(g, s)$. (This is not a significant restriction.)

Let $\tilde{u}_i(g, t)$ be a function which satisfies the conditions on $\tilde{u}_i(g, t)$ above and reduces to $u(g, s_i)$ at $t = s_i$. In (10), let $u_{\varepsilon\lambda}(t) = \tilde{u}_i(g(x(t), t), t)$ in $I_i(\varepsilon\lambda)$. Then, under the additional Assumptions 2.4 and 2.5, Theorem 3 holds with the conditioning on $\mathscr{B}_t$ replaced by conditioning on $g(\hat{x}(t), t)$. We have not given more details on the extensions to state-dependent controls, since attempts to extend the method to a more general class of controls, whose members may be discontinuous in the state, have failed so far.

REFERENCES

[1] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems*, this Journal, 4 (1966), pp. 505–527.

[2] H. J. KUSHNER, *On the stochastic maximum principle: Fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.

[3] ———, *On the stochastic maximum principle with average constraints*, Ibid., 12 (1965), pp. 13–26.

[4] F. BRODEAU, *Contribution à l'étude du contrôle optimal stochastique*, Doctoral thesis, University of Grenoble, Grenoble, France, 1968.

[5] R. F. BAUM, *Optimal control systems with stochastic boundary conditions*, Rep. 69–23, Dept. of Industrial Engrg., University of Michigan, Ann Arbor, 1969.

[6] W. FLEMING, *Stochastic Lagrange multipliers*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967.

[7] D. D. SWORDER, *On the control of stochastic systems*, Internat. J. Control, 6 (1967), pp. 179–188.

[8] ———, *On the stochastic maximum principle*, J. Math. Anal. Appl., 24 (1968), pp. 627–640.

[9] H. HALKIN, *A maximum principle of the Pontryagin type for systems described by nonlinear difference equations*, this Journal, 4 (1966), pp. 90–111.

[10] J. M. HOLTZMAN, *On the maximum principle for nonlinear discrete time systems*, IEEE Trans. Automatic Control, 4 (1966), pp. 528–547.

[11] H. J. KUSHNER, *Necessary conditions for discrete parameter stochastic otpimization problems*, Rep. 70–3, Brown University Center for Dynamical Systems; Proc. 6th Berkeley Symposium on Mathematical Statistics and Probability, to appear.

[12] ———, *Near optimal control in the presence of small stochastic perturbations*, J. Basic Engrg., 87 (1965), pp. 103–108.

[13] H. J. KUSHNER AND A. I. KLEINMAN, *Mathematical programming and the control of Markov chains*, Internat. J. Control, 13 (1971), pp. 801–820.

[14] J. NEVEU, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco, 1965.

[15] W. FLEMING AND M. NISIO, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.

[16] M. D. CANON, C. D. CULLUM, JR. AND E. POLAK, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.

SIAM J. CONTROL
Vol. 10, No. 3, August 1972

ERRATUM: EQUILIBRIUM FEEDBACK CONTROL IN LINEAR GAMES WITH QUADRATIC COSTS*

D. L. LUKES†

The second line preceding (1.2′) was omitted by the printer and reads:

"... computes a cost in the game by means of a prescribed cost ..."

* This Journal, 9 (1971), pp. 234–252. Received by the editors August 20, 1971.
† Department of Applied Mathematics, University of Virginia, Charlottesville, Virginia 22901.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

CONTROL OF FUNCTIONAL DIFFERENTIAL EQUATIONS OF RETARDED AND NEUTRAL TYPE TO TARGET SETS IN FUNCTION SPACE*

H. T. BANKS† AND G. A. KENT‡

Abstract. Optimal control of systems governed by functional differential equations of retarded and neutral type is considered. Problems with function space initial and terminal manifolds are investigated. Existence of optimal controls, regularity, and bang-bang properties are discussed. Necessary and sufficient conditions are derived and several solved examples which illustrate the theory are presented.

1. Introduction. A large number of papers have been written on the control of functional differential equations to target sets in R^n (for partial bibliographies see [4], [5], [44]). In this paper we treat a number of aspects concerning control to targets in function space. To the authors' knowledge only in the recent investigations reported in [36], [38], [39], [40], [49], [54] have others reported results for such problems. Popov [49] and Weiss [54] investigate controllability while the work of Jacobs and Kao [36], [38] concerns necessary and sufficient conditions for retarded systems. As is pointed out in § 5 below, their methods are quite different from those employed here. We treat control problems involving a fairly general class of neutral functional differential systems; a class which includes as special cases almost all nonlinear retarded systems that are of interest. Control of neutral systems in a somewhat different formulation has been investigated by Kamenskii and Khvilon [37] who use the methods of Pontryagin et al. [48] to derive necessary conditions for problems with targets sets in R^n. In § 5 we utilize the methods of Neustadt [46], [47] and Gamkrelidze [21] to obtain necessary conditions for the general nonlinear problem formulated in § 2. We also show there that under certain convexity assumptions these conditions are sufficient for normal problems with linear-in-the-state systems.

In addition to the problem formulation, § 2 also contains a motivating example which shows that boundary control problems for certain hyperbolic systems can be transformed to problems involving control of neutral functional differential equations to function space targets. Section 3 contains some of the theory (existence, representation, etc.) of neutral systems which has been developed, for the formulation used in this paper, mainly by Hale and his students and colleagues [24], [25], [26], [30]. In § 4 we present existence results for a large class of linear-in-the-state control problems. The related questions of smoothness (regularity) of controls and bang-bang properties (or lack thereof) are discussed. The paper is concluded with

* Received by the editors August 3, 1971, and in revised form October 12, 1971.

† Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported in part by the National Aeronautics and Space Administration under Grant NGL 40-002-015, in part by the Air Force Office of Scientific Research under Grant AF-AFOSR 693-67B and by the United States Army-Durham under Grant DA-31-124-ARO-D-270.

‡ Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported by the United States Navy under the Junior Line Officer Advanced Educational Program.

a section containing two solved examples along with comments concerning methods for solving examples via problem reformulation. More general necessary conditions plus a number of other solved examples may be found in the thesis of Kent [39] (see also [40]) on which much of the work reported here is based.

The following notational conventions will be adopted throughout the paper. We denote by $\mathscr{C}^n[a, b]$ the space $C([a, b], R^n)$ of R^n-valued continuous functions with the usual sup topology and by $L_p[a, b]$ the usual spaces of functions f (equivalence classes) with $|f|^p$ integrable in the sense of Lebesgue. We shall not use different symbols for various norms but let $|\cdot|$ represent the norm in whatever space may be appropriate. For example, if $x \in \mathscr{C}^n[a, b]$, $|x|$ is the sup norm of x while $|x(t)|$ represents the R^n norm of $x(t)$. The symbol $\mathscr{L}_{n\times p}$ will denote the vector space of $n \times p$ real matrices and E_p will be used for the identity in the space $\mathscr{L}_{p\times p}$. $BV[a, b]$ will represent the space of functions of bounded variation on $[a, b]$ with norm $|g| = \operatorname{Var}([a, b]; g) + |g(b)|$, where $\operatorname{Var}([a, b], g)$ is the total variation of g on $[a, b]$.

If $x:[a - h, b] \to R^n$ for $t \in [a, b]$, we denote by x_t the elements of $\mathscr{C}^n[-h, 0]$ given by $x_t(\theta) = x(t + \theta)$, $\theta \in [-h, 0]$. For systems involving hereditary dependence we shall use the notation $f(x(\cdot), t)$ to mean that $f:\mathscr{C}^n[t_0 - h, t_1] \times [t_0, t_1] \to R^n$ may depend on any or all of the values $x(s)$, $t_0 - h \leqq s \leqq t$, where $t \in [t_0, t_1]$. Examples of such dependence are $f(x(\cdot), t) = G(x(t), x(t - h), t)$, $f(x(\cdot), t) = G(x_t, t)$, $f(x(\cdot), t) = \int_{t_0-h}^{t} a(t, s)G(x(s), t)\,ds$ (see [2], [24]).

Unless it is otherwise explicitly stated, all statements involving the concept of measure will be interpreted with respect to Lebesgue measure. All integrals will be Lebesgue or Lebesgue–Stieltjes integrals [15]. Finally we shall never distinguish between a vector and its transpose since in any vector-matrix operations it will be clear what is meant.

2. Problem formulation. Let $\mathscr{T}_0$, $\mathscr{T}_1$ be given subsets of $\mathscr{C}^n[-h, 0]$ and suppose U is a specified nonempty subset of R^m. Define $\mathscr{U} = \{u:[t_0, t_1] \to R^m | u \text{ is bounded, measurable with } u(t) \in U \text{ for } t \in [t_0, t_1]\}$. We shall consider the general problem of minimizing $J = \int_{t_0}^{t_1} f^0(x(t), u(t), t)\,dt$ subject to

$$\frac{d}{dt}D(x(\cdot), t) = f(x(\cdot), u(t), t), \quad t \in [t_0, t_1], \tag{2.1}$$

$$x_{t_0} \in \mathscr{T}_0, \quad x_{t_1} \in \mathscr{T}_1, \quad u \in \mathscr{U},$$

where the function D is defined by

$$D(x(\cdot), t) = x(t) - \int_{t_0-h}^{t} d_s\mu(t, s)x(s). \tag{2.2}$$

With $\mu \not\equiv 0$ and the hypotheses specified in § 3 below, the system (2.1) is a functional differential equation (FDE) of neutral type. If $\mu \equiv 0$, the system is an FDE of retarded type. Simple examples of the type under consideration here are the differential difference equations

$$\dot{x}(t) - A(t)\dot{x}(t - h) = B(t)x(t) + C(t)x(t - h) + k(u(t), t) \tag{2.3}$$

and

$$\dot{x}(t) = B(t)x(t) + C(t)x(t - h) + k(u(t), t). \tag{2.4}$$

Many of the results obtained below can be extended to include certain types of systems involving a hereditary dependence on the control u in addition to the state x (see Kent [39]) but we shall not pursue that aspect of the problem in this paper.

There are a number of physical situations which motivate the problem as formulated above, although we shall cite only two of these here. Perhaps the simplest example where one desires to specify a terminal target in $\mathscr{C}^n[-h, 0]$ involves systems such as (2.3), (2.4). It has been recognized for many years that the true "state" for such systems is x_t, not $x(t)$. If $x(t)$ represents some error which one wishes to be driven to zero (and held there if possible) and if the error is described by (2.3) or (2.4), then it is obvious that the desired terminal condition is $x_{t_1} = 0$.

A second motivational example which we shall only sketch here involves boundary control of linear hyperbolic partial differential equations. Suppose we are given the wave equation for $w(t, x)$,

$$w_{tt} - c^2 w_{xx} = 0, \qquad t \in [0, T], \quad x \in [0, 1], \tag{2.5}$$

with boundary conditions

$$\begin{aligned} A_0(t)w_t(t, 0) + B_0(t)w_x(t, 0) &= g_0(t, w(t, 0)), \\ A_1(t)w_t(t, 1) + B_1(t)w_x(t, 1) &= g_1(t, w(t, 1)) \end{aligned} \tag{2.6}$$

and initial-terminal conditions

$$\begin{aligned} w(0, x) &= \alpha_0(x), \qquad w_t(0, x) = \alpha_1(x), \\ w(T, x) &= \beta_0(x), \qquad w_t(T, x) = \beta_1(x). \end{aligned} \tag{2.7}$$

The terms g_0, g_1 in (2.6) contain the controls for the system and hence are to be chosen from a prespecified class while α_i, β_i, $i = 0, 1$, represent given data. Suppose that A_i, B_i are continuously differentiable, g_i are absolutely continuous in t, continuously differentiable in w with g_{iw} being dominated by L_2 functions, $i = 1, 2$. In addition, assume that $\alpha_0', \alpha_1, \beta_0', \beta_1$ are absolutely continuous with L_2 derivatives ($' = d/dx$). Under the additional hypotheses $A_0(t) - B_0(t)/c \neq 0$, $A_1(t) + B_1(t)/c \neq 0$ for $t \in [0, T]$, one can derive an equivalent neutral system in the following way. Assume a solution of the form (D'Alambert)

$$w(t, x) = \varphi(t + x/c) + \psi(t - x/c).$$

Upon substitution in (2.6), followed by differentiation with respect to t and a few algebraic manipulations, one obtains a neutral system in $(\varphi', \psi') = (y, z)$ of the form

$$\begin{aligned} \dot{y}(t) + R(t)\dot{z}(t - 2/c) &= H_1(t, y(\cdot), z(\cdot)), \\ \dot{z}(t) + S(t)\dot{z}(t - 2/c) &= H_2(t, y(\cdot), z(\cdot)). \end{aligned} \tag{2.8}$$

The data given in (2.7) can be used to produce initial and terminal data in terms of (y, z) for the system (2.8). Appropriate assumptions on the boundary terms g_0, g_1

(which contain the controls for the problem) lead to a controlled system involving (2.8) for $t \in [1/c, T]$ with initial and terminal values of y specified on $[0, 1/c]$ and at $t = T$ and the corresponding values of z given on $[-1/c, 1/c]$ and $[T - 2/c, T]$. The terms H_1, H_2 are such that this initial data is sufficient to solve (2.8) for absolutely continuous (φ', ψ') having L_2 derivatives. It is not difficult to argue that this (φ, ψ) used in the D'Alambert solution above yields a solution to the original equation (2.5) in the (nonclassical) sense that $w(t, x) = \varphi(t + x/c) + \psi(t - x/c)$ is continuously differentiable with w_t, w_x being absolutely continuous and possessing L_2 partials satisfying (2.5) a.e.

The boundary conditions (2.6) include as special cases the usual boundary conditions [13], [14], [53] associated with (2.5) for transverse vibrations of a string or longitudinal vibrations in an elastic rod with elastically supported ends.

Other authors have pointed out connections between the study of hyperbolic systems and neutral FDE's. Brayton [9] and Slemrod [52] were concerned with systems arising from the study of lossless transmission lines while Cooke and Krumme [12] discussed a general method for reducing linear hyperbolic systems with nonlinear initial-boundary conditions to functional differential systems of neutral type.

3. Representation results for linear neutral systems. In this section we shall present properties of solutions of neutral systems which will be needed in the ensuing discussions. Our main results pertain to the variation of parameters representation for solutions to general linear systems. Referring to the function D defined in (2.2), we make the following standing assumptions on $\mu: R^2 \to \mathscr{L}_{n\times n}$.

Assumption 3.1. $\mu(\sigma, \theta) = 0$ for $\theta \geqq \sigma$, $\mu(\sigma, \theta) = \mu(\sigma, t_0 - h)$ for $\theta < t_0 - h$; μ is Borel measurable, continuous from the right in its first argument and continuous from the left in its second argument; $\theta \to \mu(\sigma, \theta)$ is of bounded variation on every finite θ interval, uniformly in σ; and the mapping $t \to \Gamma(\varphi, t) \equiv \int_{t_0-h}^{t} d_s\mu(t, s)\varphi(s)$ is continuous on $[t_0, t_1]$ for each fixed $\varphi \in \mathscr{C}^n[t_0 - h, t_1]$, which obviously implies that $(\varphi, t) \to \Gamma(\varphi, t)$ is continuous.

Assumption 3.2. There is a continuous nondecreasing function δ with $\delta(0) = 0$ such that for each $t \in R^1$ and $\varepsilon > 0$ we have $\operatorname{Var}([t - \varepsilon, t]; \mu(t, \cdot)) \leqq \delta(\varepsilon)$.

Specific conditions on μ directly, for which the last hypothesis in Assumption 3.1 obtains, have been given by Kent [39]. Included as a special case of these is the situation where $\mu(s, \theta) = \mathscr{J}(s, \theta) + \mathscr{A}(s, \theta)$, where $\theta \to \mathscr{J}(s, \theta)$ is a "well-behaved" jump function and $\theta \to \mathscr{A}(s, \theta)$ represents the absolutely continuous part of $\theta \to \mu(s, \theta)$. We shall not present the exact technical assumptions on $\mathscr{J}, \mathscr{A}$ here, but refer the interested reader to [39]. It suffices to remark that systems encountered in applications almost always satisfy these conditions.

We next consider solutions to

$$\begin{aligned} \frac{d}{dt} D(x(\cdot), t) &= \int_{t_0-h}^{t} d_s\eta(t, s)x(s) + g(t), \quad t \in [t_0, t_1], \\ x_{t_0} &= \varphi, \end{aligned} \tag{3.1}$$

where by a solution x we shall mean an $x \in \mathscr{C}^n[t_0 - h, t_1]$ such that $t \to D(x(\cdot), t)$ is absolutely continuous on $[t_0, t_1]$ with (3.1) being satisfied a.e. The nonhomo-

geneous term $g:[t_0, t_1] \to R^n$ will always satisfy $g \in L_1$ and we make the following hypotheses on $\eta: R^2 \to \mathscr{L}_{n\times n}$.

Assumption 3.3. $\eta(\sigma, \theta) = 0$ for $\theta \geqq \sigma$, $\eta(\sigma, \theta) = \eta(\sigma, t_0 - h)$ for $\theta < t_0 - h$; η is measurable, continuous from the left in its second variable on $(-\infty, \sigma)$; $\theta \to \eta(\sigma, \theta)$ is of bounded variation on every finite θ interval and there is an $m \in L_1^{\text{loc}}$ such that $\operatorname{Var}([t_0 - h, \sigma]; \eta(\sigma, \cdot)) \leqq m(\sigma)$.

Under the above hypotheses and Carathéordory type assumptions on f, one can prove the usual local existence and continuation theorems for solutions to (2.1) with $x_{t_0} = \varphi$, $\varphi \in \mathscr{C}^n[-h, 0]$. In addition, one can establish that (3.1) possesses a unique solution [25], [26], [30], [39]. We turn next to the "adjoint" system to (3.1) with $g \equiv 0$.

THEOREM 3.1. *Under Assumptions* 3.1, 3.2, 3.3, *for each fixed* $t \in [t_0, t_1]$ *the system*

$$\begin{aligned} &Y(s,t) = E_n + \int_s^{t^+} d_\alpha Y(\alpha, t)\mu(\alpha, s) - \int_s^t Y(\alpha, t)\eta(\alpha, s)\, d\alpha, \qquad s \in [t_0, t), \\ &Y(t,t) = E_n, \qquad Y(s,t) = 0 \quad \textit{for} \quad s > t \end{aligned} \tag{3.2}$$

has a unique solution on $[t_0, t_1]$. *This solution* $Y(s,t) \in \mathscr{L}_{n\times n}$ *is left continuous in its first argument and* $|Y(s,t)| \leqq \mathscr{B}$, $\operatorname{Var}([t_0, t_1]; Y(\cdot, t)) \leqq \mathscr{B}$ *for* $(s,t) \in [t_0, t_1] \times [t_0, t_1]$, *where* $\mathscr{B}$ *is finite and independent of* (s,t).

Proof. We assume for ease in notation (and without loss of generality) that $t_0 = 0$. The proof of existence of a unique solution and left continuity in its first argument is due to Henry [30]. We shall here only sketch the arguments, indicating how one obtains the bound $\mathscr{B}$. We note that it suffices to prove the uniform bound on the variation of $Y(\cdot, t)$ since for $s \in [t_0, t_1]$,

$$\begin{aligned} \operatorname{Var}([t_0, t_1]; Y(\cdot, t)) &\geqq |Y(t_0, t) - Y(s,t)| + |Y(s,t) - Y(t_1, t)| \\ &\geqq |Y(s,t)| - |Y(t_1, t)| \geqq |Y(s,t)| - |E_n|. \end{aligned} \tag{3.3}$$

In the proof sketched here, one actually obtains existence of the solution to (3.2) on $|s| \leqq t_1, |t| \leqq t_1$. Let $\varepsilon > 0$ be chosen sufficiently small that

$$\delta(\varepsilon) + \int_{t-\varepsilon}^t m(\theta)\, d\theta \leqq \lambda < 1 \tag{3.4}$$

for all $|t| \leqq 2t_1$, where δ is the function guaranteed in Assumption 3.2. We make the induction hypothesis (clearly true for $p = 0$) that for $|t| \leqq t_1$ the solution $Y(s,t)$ of (3.2) exists for $s \in [t - p\varepsilon, t_1]$ and satisfies $\operatorname{Var}([t - p\varepsilon, t_1]; Y(\cdot, t)) \leqq K_p$, where K_p is independent of t. We then define successive approximants by

$$Y^0(s,t) = \begin{cases} Y(s,t) & \text{if } s \in (t - p\varepsilon, t_1], \\ Y(t - p\varepsilon, t) & \text{if } s \in [-t_1, t - p\varepsilon] \end{cases} \tag{3.5}$$

and

$$Y^k(s,t) = \begin{cases} Y(s,t) \quad \text{if } s \in (t - p\varepsilon, t_1], \\ E_n + \displaystyle\int_s^{t^+} d_\alpha Y^{k-1}(\alpha, t)\mu(\alpha, s) - \int_s^t Y^{k-1}(\alpha, t)\eta(\alpha, s)\, d\alpha \\ \qquad\qquad \text{if } s \in [-t_1, t - p\varepsilon] \end{cases} \tag{3.6}$$

for $k = 1, 2, \cdots$. Using (3.4) and these definitions along with Assumptions 3.2, 3.3, one can easily show that

$$\| Y^{k+1}(\cdot, t) - Y^k(\cdot, t)\| \leqq \lambda \| Y^k(\cdot, t) - Y^{k-1}(\cdot, t)\|$$

and hence

$$\| Y^{k+1}(\cdot, t) - Y^k(\cdot, t)\| \leqq \lambda^k \| Y^1(\cdot, t) - Y^0(\cdot, t)\|,$$

where $\|g\| \equiv \text{Var}([t - (p+1)\varepsilon, t - p\varepsilon]; g)$. Observing that one actually has $Y^k(s, t) = Y(s, t)$ for $s \in [t - p\varepsilon, t_1]$, one sees that $Y^k(\cdot, t)$ converges in $BV[t - (p+1)\varepsilon, t - p\varepsilon]$ to a function $\tilde{Y}(\cdot, t)$. Letting $k \to \infty$ in the above approximants we see that this extends the solution from $[t - p\varepsilon, t_1]$ to $[t - (p+1)\varepsilon, t_1]$. A finite number of induction steps on p yields existence as claimed. The left continuity and uniqueness follow directly from the hypotheses on μ, η and the equation for Y.

Returning to the arguments involving $Y^k(\cdot, t)$, we have that

$$\begin{aligned}\| Y^k(\cdot, t)\| &\leqq \| Y^k(\cdot, t) - Y^{k-1}(\cdot, t)\| + \| Y^{k-1}(\cdot, t)\| \\ &\leqq \{\lambda^{k-1} + \lambda^{k-2} + \cdots + \lambda^1 + 1\} \| Y^1(\cdot, t) - Y^0(\cdot, t)\| + \| Y^0(\cdot, t)\| \\ &\leqq \frac{1}{1-\lambda} \| Y^1(\cdot, t) - Y^0(\cdot, t)\| + \| Y^0(\cdot, t)\|.\end{aligned}$$

But $\| Y^0(\cdot, t)\| = 0$ and using elementary arguments with the hypotheses on η, μ, the definitions of Y^0, Y^1, one can easily show that

$$\| Y^1(\cdot, t) - Y^0(\cdot, t)\| \leqq K_p \left\{ M + 2 \int_{t-(p+1)\varepsilon}^{t-p\varepsilon} m(\theta)\, d\theta \right\}$$

for $|t| \leqq t_1$. It follows that

$$\| Y^k(\cdot, t)\| \leqq B_p,$$

where B_p is independent of $t, |t| \leqq t_1$. Since $\| Y^k(\cdot, t) - \tilde{Y}(\cdot, t)\| \to 0$, we obtain $\| \tilde{Y}(\cdot, t)\| \leqq B_p$. Thus

$$\begin{aligned}\text{Var}([t - (p+1)\varepsilon, t_1]; Y(\cdot, t)) &\leqq \| \tilde{Y}(\cdot, t)\| + \text{Var}([t - p\varepsilon, t_1]; Y(\cdot, t)) \\ &\leqq B_p + K_p = K_{p+1},\end{aligned}$$

where K_{p+1} is independent of $t, |t| \leqq t_1$. The finite number of induction steps on p then produce the bound $\mathscr{B}$ independent of t.

THEOREM 3.2. *Let x be the solution of* (3.1) *under Assumptions* 3.1, 3.2, 3.3. *Then for $t \in [t_0, t_1]$,*

$$x(t) = Y(t_0, t)D(\varphi, t_0) + \int_{t_0-h}^{t_0^-} d_\beta \gamma(t, \beta)\varphi(\beta) + \int_{t_0}^{t} Y(\beta, t)g(\beta)\, d\beta, \tag{3.7}$$

where Y is given by (3.2) *and*

$$\gamma(t, \beta) \equiv -\int_{t_0}^{t^+} d_\alpha Y(\alpha, t)\mu(\alpha, \beta) + \int_{t_0}^{t} Y(\alpha, t)\eta(\alpha, \beta)\, d\alpha.$$

The proof of this theorem is due to Henry [30]. We shall omit it here since it involves a standard type of argument making use of integration by parts, an

unsymmetric Fubini theorem [10], and the equation for Y. We note that for $\mu \equiv 0$ the adjoint system (3.2) and the representation (3.7) reduce to that for retarded systems [2], [3].

Remark 3.1. We make some further comments about the solution Y of (3.2) which may correctly be regarded as a "fundamental matrix solution." It is not difficult to show [39] that for fixed s, the function $t \to Y(s, t)$ (which is, in general, discontinuous) is in BV and satisfies

$$Y(s,t) = E_n + \int_s^t d_\theta \mu(t,\theta) Y(s,\theta) + \int_s^t d\beta \int_s^\beta d_\theta \eta(\beta,\theta) Y(s,\theta)$$

for $t \geqq s$, with $Y(s, t) = 0$ for $t < s$. Note that this is just the integrated form of (3.1) with $g \equiv 0$. In a subsequent section (§ 5) of this paper, it will be essential that the mapping $(s, t) \to Y(s, t)$ be Borel measurable. (This is needed in order to use Y as the measure in the unsymmetric Fubini theorem [10].) This will be true under varied assumptions on μ. For example, Kent [39] has shown that it is sufficient that $t \to \mu(t, \theta)$ be of bounded variation on each finite t interval for each fixed θ. Henry [30] has established Borel measurability of Y under other assumptions. Since we do not wish to become involved in these technical details here and since any system of interest to us in this paper would meet either Kent's or Henry's assumptions, we make the standing hypothesis that $(s, t) \to Y(s, t)$ is Borel measurable.

Before presenting the final results of this section we must make the following definition. For $\Phi \subset \mathscr{C}^n[-h, 0]$, $K \in L_1[t_0, t_1]$ and $X \subset R^n$, define $C([t_0 - h, t_1], X; \Phi, K) \subset \mathscr{C}^n[t_0 - h, t_1]$ by $C([t_0 - h, t_1], X; \Phi, K) = \{x \in \mathscr{C}^n[t_0 - h, t_1] | x_{t_0} \in \Phi, x(t) \in X \text{ for } t \in [t_0 - h, t_1], |(d/dt)D(x(\cdot), t)| \leqq K(t) \text{ a.e. } t \in [t_0, t_1]\}$.

THEOREM 3.3. *In addition to Assumptions* 3.1 *and* 3.2 *assume that* μ *has the following property: there exist* $l > 0, L > 0$ *such that*

$$\int_{t-l}^t |d_\theta\{\mu(t,\theta) - \mu(s,\theta)\}| \leqq L|t - s| \quad \text{for } s \leqq t. \tag{3.8}$$

Then X, Φ *being compact implies* $C([t_0 - h, t_1], X; \Phi, K)$ *is a compact subset of* $\mathscr{C}^n[t_0 - h, t_1]$. *It is also convex if* X, Φ *are.*

Proof. Convexity follows from the linearity of $D(x(\cdot), t) = x(t) - \Gamma(x, t)$. For any $\varphi \in \Phi$, we define x_φ by $x_\varphi(t) = \varphi(t - t_0), t \in [t_0 - h, t_0], x_\varphi(t) = \varphi(0), t > t_0$. Then for $x \in C([t_0 - h, t_1], X; \Phi, K)$ with $x_{t_0} = \varphi$ we have

$$\begin{aligned} |x(t) - x(t_0)| &\leqq |D(x(\cdot), t) - D(x(\cdot), t_0)| + |\Gamma(x, t) - \Gamma(x_\varphi, t)| \\ &\quad + |\Gamma(x_\varphi, t) - \Gamma(x_\varphi, t_0)| \\ &\leqq \int_{t_0}^t K(s)\, ds + \left| \int_{t_0}^t d_\theta \mu(t,\theta)\{x(\theta) - x(t_0)\} \right| + |\Gamma(x_\varphi, t) - \Gamma(x_\varphi, t_0)|. \end{aligned}$$

Choosing p such that $0 < p \leqq l$ and $\delta(p) < 1$, we thus find for $t \in [t_0, t_0 + p]$ that

$$\begin{aligned} |x(t) - x(t_0)| \leqq \int_{t_0}^t K(s)\, ds + \delta(p)\|x - x_\varphi\|_t \\ + \sup\{|\Gamma(x_\psi, t) - \Gamma(x_\psi, t_0)| : t \in [t_0, t_0 + p], \psi \in \Phi\}, \end{aligned}$$

where $\|x\|_t = \sup\{|x(s)| : s \in [t_0 - h, t]\}$. But since the right side of this expression

is nondecreasing in t, we obtain

$$\|x - x_\varphi\|_t \leqq \int_{t_0}^{t} K(s)\,ds + \delta(p)\|x - x_\varphi\|_t + \sup_{t,\psi} |\Gamma(x_\psi, t) - \Gamma(x_\psi, t_0)|$$

or

$$\|x - x_\varphi\|_t \leqq b\left\{\int_{t_0}^{t_0+p} K(s)\,ds + \sup_{t,\psi} |\Gamma(x_\psi, t) - \Gamma(x_\psi, t_0)|\right\} = R,$$

where $b = 1/(1 - \delta(p))$. We remark that the sup term is finite since Γ is continuous, Φ is compact. For $t_0 \leqq \tau \leqq t \leqq t_0 + p$ we therefore have

$$\begin{aligned}
|x(t) - x(\tau)| &\leqq |D(x(\cdot), t) - D(x(\cdot), \tau)| + |\Gamma(x, t) - \Gamma(x, \tau)| \\
&\leqq \int_{\tau}^{t} K(s)\,ds + |\Gamma(x_\varphi, t) - \Gamma(x_\varphi, \tau)| \\
&\quad + \left| \int_{t_0}^{t} d_s[\mu(t, s) - \mu(\tau, s)]\{x(s) - x_\varphi(s)\} \right| \\
&\leqq \int_{\tau}^{t} K(s)\,ds + \sup\{|\Gamma(x_\psi, t) - \Gamma(x_\psi, \tau)| : \psi \in \Phi\} + L|t - \tau|R.
\end{aligned}$$

From the continuity of Γ and the compactness of Φ, it follows that the elements of $[C([t_0 - h, t_1], X; \Phi, K)$ form an equicontinuous family on $[(t_0 - h, t_0 + p]$. Since the restrictions of these elements to $[t_0 - h, t_0 + p]$ constitute a bounded subset of $\mathscr{C}^n[t_0 - h, t_0 + p]$, use of the Arzela–Ascoli theorem here, followed by repetition of the above arguments on $[t_0 + p, t_0 + 2p], [t_0 + 2p, t_0 + 3p], \cdots$ (for a finite number of steps), leads to the conclusion that $C([t_0 - h, t_1], X; \Phi, K)$ is a conditionally compact subset of $\mathscr{C}^n[t_0 - h, t_1]$. Using the compactness of X, Φ and the continuity of D, it is not difficult to argue that $C([t_0 - h, t_1], X; \Phi, K)$ is a closed, hence compact, subset of $\mathscr{C}^n[t_0 - h, t_1]$.

4. Existence, regularity and bang-bang results. We shall consider first systems that are linear in the state, i.e., system (3.1) with $g(t) = k(u(t), t)$. We shall approach the questions of existence, smoothness of controls and bang-bang properties by considering attainable sets in $\mathscr{C}^n[-h, 0]$. We make the following assumption.

Assumption 4.1. The mapping $k: R^m \times R^1 \to R^n$ is continuous, the set $U \subset R^m$ is compact, and $k(U, t) \equiv \{k(u, t) | u \in U\}$ is convex for each t.

Define the family $\mathfrak{F}$ by

$$\begin{aligned}
\mathfrak{F} = \Big\{ \mathfrak{f} : \mathscr{C}^n[t_0 - h, t_1] \times [t_0, t_1] \to R^n \,|\, \mathfrak{f}(x(\cdot), t) \\
= \int_{t_0 - h}^{t} d_s\eta(t, s)x(s) + k(u(t), t), u \in \mathscr{U} \Big\},
\end{aligned} \tag{4.1}$$

where $\mathscr{U}$ is the class of admissible controls as defined in § 2. We are thus considering

$$\begin{aligned}
\frac{d}{dt} D(x(\cdot), t) &= \mathfrak{f}(x(\cdot), t), \qquad t \in [t_0, t_1], \\
x_{t_0} &= \varphi
\end{aligned} \tag{4.2}$$

for $(\varphi, \mathfrak{f}) \in \Phi \times \mathfrak{F}$. We assume of course throughout that μ, η satisfy Assumptions 3.1, 3.2, 3.3. Recalling that $\|y\|_t = \sup\{|y(s)| : s \in [t_0 - h, t]\}$ we then have the following lemma.

LEMMA 4.1. *In addition to the above hypotheses, assume Φ is a bounded subset of $\mathscr{C}^n[-h, 0]$. Then there is an $M > 0$ such that $\|x(\varphi, \mathfrak{f})\|_t \leqq M$ for $t \in [t_0, t_1]$, $(\varphi, \mathfrak{f}) \in \Phi \times \mathfrak{F}$, where $x(\varphi, \mathfrak{f})$ denotes the solution to* (4.2) *for $(\varphi, \mathfrak{f}) \in \Phi \times \mathfrak{F}$.*

Proof. It is easy to see that there is a constant $\tilde{d}$ and an L_1 function $\tilde{m}$ such that $|\mathfrak{f}(x(\cdot), t)| \leqq \tilde{m}(t)[\|x\|_t + \tilde{d}]$ for every $\mathfrak{f} \in \mathfrak{F}$. Letting $x = x(\varphi, \mathfrak{f})$, $(\varphi, \mathfrak{f}) \in \Phi \times \mathfrak{F}$, we have

$$x(t) = x(t_0) + \int_{t_0}^{t} d_s\mu(t, s)x(s) + \int_{t_0-h}^{t_0} d_s[\mu(t, s) - \mu(t_0, s)]x(s) + \int_{t_0}^{t} \mathfrak{f}(x(\cdot), s)\, ds.$$

Choosing p such that $0 < \delta(p) < 1$ and letting $b = 1/(1 - \delta(p))$, we find for $t \in [t_0, t_0 + p]$ that

$$\|x\|_t \leqq |\varphi| + \delta(p)\|x\|_t + 2\delta(t_1 - t_0 + h)|\varphi| + \int_{t_0}^{t} \tilde{m}(s)[\|x\|_s + \tilde{d}]\, ds.$$

It follows that

$$\tilde{d} + \|x\|_t \leqq b\left\{[1 + 2\delta(t_1 - t_0 + h)][|\varphi| + \tilde{d}] + \int_{t_0}^{t} \tilde{m}(s)[\|x\|_s + \tilde{d}]\, ds\right\}.$$

An application of Gronwall's inequality on $[t_0, t_0 + p]$ followed by a repetition of the above arguments on $[t_0 + p, t_0 + 2p], \cdots, [t_0 + (N - 1)p, t_1]$ yields the desired result in view of the boundedness assumption on Φ.

LEMMA 4.2. *Under Assumptions* 3.1, 3.2, 3.3, 3.8 *and Assumption* 4.1, *Φ being compact in $\mathscr{C}^n[-h, 0]$ implies $\mathscr{A} \equiv \{x(\varphi, \mathfrak{f}) | (\varphi, \mathfrak{f}) \in \Phi \times \mathfrak{F}\}$ is an equicontinuous subset of $\mathscr{C}^n[t_0 - h, t_1]$.*

Proof. From Lemma 4.1 it follows that for any $x \in \mathscr{A}$,

$$\frac{d}{dt} D(x(\cdot), t) \leqq \tilde{m}(t)[\|x\|_t + \tilde{d}] \leqq \tilde{m}(t)[M + \tilde{d}] = K(t)$$

on $[t_0, t_1]$, where $K \in L_1$. The same lemma guarantees existence of a compact $X \subset R^n$ such that $x(t) \in X$, $t \in [t_0 - h, t_1]$, $x \in \mathscr{A}$. Thus $\mathscr{A}$ is a subset of $C([t_0 - h, t_1], X; \Phi, K)$ which, by Theorem 3.3, is a compact subset of $\mathscr{C}^n[t_0 - h, t_1]$. The equicontinuity of $\mathscr{A}$ thus follows from a well-known theorem [15, p. 266].

We define, for each $t \in [t_0, t_1]$, the attainable set at time t given by

$$\mathscr{A}_t(\Phi, \mathfrak{F}) = \{z \in \mathscr{C}^n[-h, 0] | z = x_t(\varphi, \mathfrak{f}), (\varphi, \mathfrak{f}) \in \Phi \times \mathfrak{F}\}.$$

Using the representation results given in Theorem 3.2 for the solutions $x(\varphi, \mathfrak{f})$, we can write

$$\mathscr{A}_t(\Phi, \mathfrak{F}) = \mathscr{R}_t(\Phi, 0) + \mathscr{R}_t(0, \mathscr{U}),$$

where

$$\mathscr{R}_t(\Phi, 0) \equiv \left\{ y_t \in \mathscr{C}^n[-h, 0] \,|\, y(s) = Y(t_0, s)D(\varphi, t_0) + \int_{t_0-h}^{t\bar{0}} d_\beta \gamma(s, \beta)\varphi(\beta) \right.$$

$$\left. \text{for } s > t_0, y(s) = \varphi(s - t_0) \text{ for } s \in [t_0 - h, t_0], \varphi \in \Phi \right\}$$

and

$$\mathscr{R}_t(0, \mathscr{U}) \equiv \left\{ y_t \in \mathscr{C}^n[-h, 0] \,|\, y(s) = \int_{t_0}^{s} Y(\beta, s)k(u(\beta), \beta)\, d\beta \text{ for } s > t_0, \right.$$

$$\left. y(s) = 0 \text{ for } s \in [t_0 - h, t_0], u \in \mathscr{U} \right\}.$$

We note that $\mathscr{R}_t(\Phi, 0)$ consists of restrictions of solutions to (3.1) with $g \equiv 0$ and initial data $\varphi \in \Phi$, while $\mathscr{R}_t(0, \mathscr{U})$ is the set of restrictions of solutions to (3.1) with $g(t) = k(u(t), t), u \in \mathscr{U}$, and initial data $\varphi \equiv 0$.

For fixed $t \in (t_0, t_1]$, let us first consider the set $\mathscr{R}_t(0, \mathscr{U})$. We define the set $\mathscr{K}(\mathscr{U})$ and the mapping $T^t : L_1[t_0, t_1] \to \mathscr{C}^n[-h, 0]$ by

$$\mathscr{K}(\mathscr{U}) = \{g \in L_1[t_0, t_1] \,|\, g(s) = k(u(s), s), u \in \mathscr{U}\}$$

and

$$T^t(g)(\theta) = \begin{cases} \displaystyle\int_{t_0}^{t+\theta} Y(\beta, t + \theta)g(\beta)\, d\beta & \text{for } t + \theta > t_0, \\ 0 & \text{for } t + \theta \leqq t_0 \end{cases}$$

for $\theta \in [-h, 0]$, so that $\mathscr{R}_t(0, \mathscr{U}) = T^t(\mathscr{K}(\mathscr{U}))$.

LEMMA 4.3. *Under Assumption* 4.1, $\mathscr{R}_t(0, \mathscr{U})$ *is a closed subset of* $\mathscr{C}^n[-h, 0]$.

Proof. The ideas in this proof are by now quite familiar to control theorists [32, pp. 18–23], [11], [17], [31]. First, from arguments similar to Filippov's [19], it follows easily that $\mathscr{K}(\mathscr{U})$ is a weakly sequentially closed subset of $L_1[t_0, t_1]$. From Assumption 4.1 one obtains [15, p. 292] that $\mathscr{K}(\mathscr{U})$ is weakly sequentially compact in $L_1[t_0, t_1]$ and hence by Eberlein–Smulian [15, p. 430] the weak closure of $\mathscr{K}(\mathscr{U})$ is weakly compact. But the weak closure of $\mathscr{K}(\mathscr{U})$ is the same as its weak sequential closure [15, p. 434]. Hence $\mathscr{K}(\mathscr{U})$ is a weakly compact subset of $L_1[t_0, t_1]$.

The map T^t is clearly continuous with respect to the strong topologies of L_1 and $\mathscr{C}^n[-h, 0]$, and hence continuous with respect to the weak topologies on these spaces [15, p. 422]. It follows that $T^t(\mathscr{K}(\mathscr{U}))$ is weakly compact and hence weakly (a fortiori strongly) closed in $\mathscr{C}^n[-h, 0]$.

Remark 4.1. Assumption 4.1 under which Lemma 4.3 obtains can be relaxed. Using arguments similar to those of Jacobs [35, p. 416] one can show that $\mathscr{K}(\mathscr{U})$ is weakly sequentially closed under the assumptions: $t \to U(t)$ defines an upper semicontinuous mapping with range in the collection of nonempty compact subsets of R^m; $k(U(t), t)$ is convex for each $t \in [t_0, t_1]$; and $u \to k(u, t)$ is continuous for each t, $t \to k(u, t)$ is measurable, and there is an $m \in L_1$ such that $|k(u, t)| \leqq m(t)$,

$u \in U(t)$. The other arguments in the proof of Lemma 4.3 then hold without change. Comments on relaxing the convexity assumptions will be made below.

Under the assumptions of Lemma 4.2, we see that $\mathscr{R}_t(0, \mathscr{U})$ is an equicontinuous subset of $\mathscr{C}^n[-h, 0]$. Since it is also bounded, the Arzela–Ascoli theorem implies that $\mathscr{R}_t(0, \mathscr{U})$ is conditionally compact. Lemma 4.3 then yields the compactness of $\mathscr{R}_t(0, \mathscr{U})$.

Next let us consider $\mathscr{R}_t(\Phi, 0)$. For Φ compact in $\mathscr{C}^n[-h, 0]$, to show $\mathscr{R}_t(\Phi, 0)$ is compact in $\mathscr{C}^n[-h, 0]$ it suffices to show that the mapping $\varphi \to y_t(\varphi)$ is continuous, where y_t is as given in the definition of $\mathscr{R}_t(\Phi, 0)$. Clearly, it is enough to demonstrate that for $t \geqq t_0 + h$ the mapping $\varphi \to z_t(\varphi) \in \mathscr{C}^n[-h, 0]$ is continuous, where

$$z_t(\varphi)(\theta) \equiv Y(t_0, t + \theta)D(\varphi, t_0) + \int_{t_0 - h}^{t\bar{0}} d_\beta \gamma(t + \theta, \beta)\varphi(\beta), \quad \theta \in [-h, 0].$$

We have

$$\begin{aligned} |z_t(\varphi) - z_t(\psi)| &\leqq \sup_{\theta \in [-h, 0]} |Y(t_0, t + \theta)| |D(\varphi, t_0) - D(\psi, t_0)| \\ &\quad + \sup_{\theta \in [-h, 0]} \left| \int_{t_0 - h}^{t\bar{0}} d_\beta \gamma(t + \theta, \beta)[\varphi(\beta) - \psi(\beta)] \right| \\ &\leqq \sup_{\theta} |Y(t_0, t + \theta)| |D(\varphi, t_0) - D(\psi, t_0)| \\ &\quad + \sup_{\theta} \operatorname{Var}([t_0 - h, t_0]; \gamma(t + \theta, \cdot))|\varphi - \psi|. \end{aligned}$$

From the continuity of D, it suffices to show that the sup terms are finite. Using the definition of γ and Assumptions 3.1, 3.3 on μ, η, one can show easily that for $\theta \in [-h, 0]$,

$$\begin{aligned} \operatorname{Var}([t_0 - h, t_0]; \gamma(t + \theta, \cdot)) \leqq M_1 &\int_{t_0}^{t + \theta^+} |d_\alpha Y(\alpha, t + \theta)| \\ &+ \int_{t_0}^{t + \theta} |Y(\alpha, t + \theta)| m(\alpha)\, d\alpha. \end{aligned}$$

The bounds guaranteed in Theorem 3.1 then imply that the sup terms are indeed finite.

THEOREM 4.1. *Under Assumptions* 3.1, 3.2, 3.3, 3.8 *and Assumption* 4.1, Φ *being compact in* $\mathscr{C}^n[-h, 0]$ *implies* $\mathscr{A}_t(\Phi, \mathfrak{F})$ *is compact in* $\mathscr{C}^n[-h, 0]$, $t \in [t_0, t_1]$. *Furthermore, the mapping* $t \to \mathscr{A}_t(\Phi, \mathfrak{F})$ *is continuous with respect to the Hausdorff metric* [7].

Proof. The first conclusion of the theorem is evident from the compactness of $\mathscr{R}_t(\Phi, 0)$ and $\mathscr{R}_t(0, \mathscr{U})$. The second assertion follows easily from the fact (Lemma 4.2) that $t \to x_t(\varphi, \mathfrak{f})$ is continuous uniformly in $(\varphi, \mathfrak{f}) \in \Phi \times \mathfrak{F}$.

Remark 4.2. The continuity of $t \to \mathscr{A}_t(\Phi, \mathfrak{F})$ can be proved for retarded systems ($\mu \equiv 0$) by the usual arguments ([42, pp. 70–71], [45, p. 114]) involving the variation of parameters representation (Theorem 3.2). These arguments depend very much on the continuity of $t \to Y(\alpha, t)$. For neutral systems, this continuity requirement is not met and hence a direct extension of the usual arguments is not possible. The compactness arguments can also be made somewhat more directly for the retarded case by taking advantage of the continuity of $t \to Y(\alpha, t)$.

The results given in Theorem 4.1 are sufficient to obtain existence theorems for a wide class of linear-in-the-state problems with initial and terminal manifolds in $\mathscr{C}^n[-h, 0]$. Since the proofs involve well-known arguments [32], [42], [45], we shall only list a few of these problems here. (In each case it is assumed that $\mathscr{T}_0 = \Phi$ is compact and $\mathscr{T}_1$ is closed.)

(i) The time-optimal problem for hitting a target set $\mathscr{T}_1 \subset \mathscr{C}^n[-h, 0]$, starting from an initial manifold $\mathscr{T}_0 = \Phi$. (In the formulation of § 2, take $f^0 = 1$ and t_1 not fixed.)

(ii) The problem of minimizing $P(x_{t_1})$ subject to $x_{t_i} \in \mathscr{T}_i$, $i = 0, 1$, where $P : \mathscr{C}^n[-h, 0] \to R^1$ is continuous.

(iii) Minimization of $J = \int_{t_0}^{t_1} \{A^0 x(t) + k^0(u(t), t)\}\, dt$ subject to $x_{t_i} \in \mathscr{T}_i$, $i = 0, 1$. (The usual device of augmenting the system and then minimizing $P(\hat{x}) = x^0(t_1)$ allows a direct application of Theorem 4.1 here.)

Various generalizations are possible, e.g. allowing the target $\mathscr{T}_1$ to depend on time [4], [32], [42], [45]. We shall not pursue these matters in this paper since our main interest concerns problems with $\mathscr{T}_0, \mathscr{T}_1$ each consisting of a single point in $\mathscr{C}^n[-h, 0]$.

Kent has shown that fairly general existence results such as those by Jacobs [35] can be extended to include systems described by certain nonlinear neutral equations. We shall not present those arguments here, but instead refer the reader to [39]. One type of nonlinear existence result which we should mention here is needed for a complete discussion of Example 6.2 below. Briefly, suppose that in the formulation in § 2, f of (2.1) has the form $f(x(\cdot), u(t), t) = \int_{t_0-h}^{t} d_s\eta(t, s)x(s) + B(t)u(t), f^0(x(t), u(t), t) = x(t)Qx(t) + u(t)Ru(t)$, where $Q \geqq 0$, $R > 0$ and $\mathscr{T}_0 = \{\varphi\}, \mathscr{T}_1 = \{\zeta\}$ where φ, ζ are given in $\mathscr{C}^n[-h, 0]$. Existence results for this problem follow from standard arguments [42] which we shall only sketch here. Letting $\{u_n\}$ be chosen, $u_n \in \mathscr{U}$, $x_{n_{t_0}} = \varphi$, $x_{n_{t_1}} = \zeta$, so that

$$J(u_n) = \int_{t_0}^{t_1} x_n Q x_n + u_n R u_n \to \beta$$

$$= \inf \{J(u) | u \in \mathscr{U}, x_{t_0}(u) = \varphi, x_{t_1}(u) = \zeta\},$$

one sees that $\int_{t_0}^{t_1} u_n R u_n$ is bounded, and hence $\{u_n\}$ is bounded in $L_2[t_0, t_1]$. Choosing a weakly convergent subsequence $\{u_{n_k}\}$, it is not hard to argue that the corresponding trajectories $\{x_{n_k}\}$ satisfy $x_{n_k}(t) \to x^*(t), t \in [t_0, t_1]$, where x^* is the trajectory corresponding to u^*, the weak limit of $\{u_{n_k}\}$. Use of the weak lower-semicontinuity of $u \to \int uRu$ along with other well-known arguments yields $J(u^*) \leqq \beta$. If $U = R^m$, then[1] $u^* \in \mathscr{U}$ and the proof is completed. If U is compact convex in R^m, then [15, p. 422] some sequence of convex combinations of the u_{n_k} converges in $L_2[t_0, t_1]$ to u^*. It follows that $u^*(t) \in U$ a.e. and again the proof is concluded.

We turn next to the questions of regularity (smoothness) of controls and bang-bang properties (U is taken compact in R^m). Many authors have investigated these questions for finite and certain types of infinite-dimensional systems. An often considered question [18], [27], [28], [29], [32], [42] is the following: Given a "state" that is attainable from a given initial "state" employing a measurable

[1] In this case we take $\mathscr{U}$ as $L_2[t_0, t_1]$ and do not insist that u^* be pointwise bounded as required in the formulation of § 2.

control, is it possible to attain the same "state" using a piecewise continuous bang-bang control? An affirmative answer to this question has been given by Banks and Jacobs [4] for some classes of retarded systems when the attainable "states" are taken in R^n. The methods used are extensions of ideas due to Halkin and Hendricks [27], [28], [29]. The results in [4] can be extended to include certain types of neutral systems (for example, (2.3) above when A, B, C are analytic). However, our interest here is in attainable "states" in $\mathscr{C}^n[-h, 0]$. Since it is known that there is no infinite-dimensional analogue of the Lyapunov theorem which is the basis of the above arguments [27], [28], [29], we face considerable difficulties. In fact, in light of known results for linear ODE systems (see [32, p. 111]) concerning bang-bang results for trajectories (attainable "states" in $\mathscr{C}^n[-h, 0]$ correspond to terminal segments of trajectories: $x(s), s \in [t_1 - h, t_1]$) one might suspect that the best that can be obtained is a density theorem. But what if one is willing to make assumptions on the FDE (such as A or C nonsingular in (2.3) or (2.4)) so that ODE's are not special cases of the FDE? The authors did this for scalar autonomous retarded systems, using the methods of Hale [23] involving a finite-dimensional projection argument (for which the Lyapunov-type results are valid). Certain results on eigenfunction expansions of solutions to FDE's [6] were also needed since a limiting process in a space of closed subsets of $\mathscr{C}[-h, 0]$ was employed. This led only to the density theorem one would certainly expect to be true. The following examples demonstrate the futility of our efforts.

Example 4.1. Consider the retarded system

$$\dot{x}(t) = x(t - 1) + u(t), \qquad t \in [0, 3],$$

with $U = \{v \in R^1 : |v| \leqq 1\}$ and $\zeta(t) = \frac{1}{2}(3 - t), t \in [2, 3]$. It is not hard to show that ζ can be attained by a measurable control u with $u(t) \in U$, starting from the initial function $\varphi(t) = -t, t \in [-1, 0]$. But ζ is not attainable with a bang-bang control from any initial φ. Suppose it were, with the bang-bang control denoted by u^0. Since $\dot{x}(t) = -\frac{1}{2}$ for $t \in [2, 3]$ and $|u^0(t)| = 1$, we see from the equation that $|-\frac{1}{2} - x(t - 1)| = 1$ or $x(t - 1) = \frac{1}{2}, t \in [2, 3]$. This implies $x(t) = \frac{1}{2}$ and $\dot{x}(t) = 0$ for $t \in [1, 2]$. But then $0 = x(t - 1) + u^0(t), t \in [1, 2]$, and $|u^0(t)| = 1$ implies $|x(t - 1)| = 1, t \in [1, 2]$, or $|x(t)| = 1, t \in [0, 1]$, contradicting the fact that $x(1) = \frac{1}{2}$.

Example 4.2. Consider the neutral system

$$\dot{x}(t) = \dot{x}(t - 1) + u(t), \qquad t \in [0, 2],$$

with U as in the previous example and $\zeta(t) = 2 - t, t \in [1, 2]$. The assumption that u is bang-bang and ζ is attained leads easily to the conclusion that $|\dot{\varphi}(t)| = 1$ or $\dot{\varphi}(t) = -3$, where φ is the initial function. But it is easy to demonstrate that there are initial functions φ (e.g. $\varphi(t) \equiv 1$) with $|\dot{\varphi}(t)| \neq 1, \dot{\varphi}(t) \neq -3$, and admissible controls u such that ζ is attainable from φ using u. Thus there are initial functions φ for which the attainable set in $\mathscr{C}^n[-h, 0]$ from φ using bang-bang controls is a proper subset of the set attained using all admissible controls.

The density theorem mentioned above can also be obtained under less restrictive assumptions as an easy application of a result due to Fattorini. For example, consider (3.1) with $\mu \equiv 0$, $\Phi = \{\varphi\}$, $g(t) = B(t)u(t)$, $B \in L_1$, and define $U = \prod_{i=1}^m [-1, 1]$, $U_e = \prod_{i=1}^m \{-1, 1\}$ so that U_e consists of the vertices of the "cube" U. Let $\mathscr{U} \equiv \{u : [t_0, t_1] \to R^m | u \text{ measurable}, u(t) \in U\}$ while $\mathscr{U}^{\#} \equiv \{u : [t_0, t_1]$

$\to R^m | u$ piecewise continuous, $u(t) \in U_e\}$. Then $\mathscr{A}_t(\Phi, \mathscr{U})$, $\mathscr{A}_t(\Phi, \mathscr{U}^\#)$ are defined as subsets of $\mathscr{C}^n[-h, 0]$ in the usual way using the representation results of § 3. It is not hard to argue that one could make use of Bochner integrals [33] in place of the Lebesgue integrals in defining $\mathscr{A}_t(\Phi, \mathscr{U})$, $\mathscr{A}_t(\Phi, \mathscr{U}^\#)$. Under the standing hypotheses of § 3, it is then easy to verify that Lemma 1 of [18] is applicable and thus $\mathscr{A}_t(\Phi, \mathscr{U}^\#)$ is dense in $\mathscr{A}_t(\Phi, \mathscr{U})$ in the norm of $\mathscr{C}^n[-h, 0]$.

Remark 4.3. The examples above show that while $\mathscr{A}_t(\Phi, \mathscr{U}^\#)$ may be dense in $\mathscr{A}_t(\Phi, \mathscr{U})$, it will in general be a proper subset of $\mathscr{A}_t(\Phi, \mathscr{U})$ and hence is not closed in $\mathscr{C}^n[-h, 0]$. Referring to Remark 4.1, we see that one therefore would not obtain the closure results of Lemma 4.3 if the convexity assumptions of Assumption 4.1 are relaxed. Note that this differs from the situation for linear ODE and FDE systems where one can obtain existence results in the absence of convexity assumptions when the "state" is taken in R^n [4], [35], [45].

5. Necessary and sufficient conditions. In this section we shall first derive necessary conditions for the problem given in § 2 with t_0, t_1 fixed and $\mathscr{T}_0 = \{\varphi\}$, $\mathscr{T}_1 = \{\zeta\}$, where φ, ζ are given in $\mathscr{C}^n[-h, 0]$, i.e., the fixed endpoint problem in control theory. Necessary conditions for problems with t_1 variable (including the time-optimal problem) and with more general manifolds $\mathscr{T}_0, \mathscr{T}_1$ have been given in [39], [40]. Considering the problem as formulated in § 2, we define $\hat{f} = (f^0, f)$ and in addition to the standing hypotheses of § 3 on μ we assume: $\hat{f}: \mathscr{C}^n[t_0 - h, t_1] \times U \times [t_0, t_1] \to R^{n+1}$ is continuously differentiable in x for each fixed $(u, t) \in U \times [t_0, t_1]$ and Borel measurable in (u, t) for each fixed $x \in \mathscr{C}^n[t_0 - h, t_1]$. Furthermore, given any $u \in \mathscr{U}$ and compact convex $X \subset R^n$, there is an $m \in L_1[t_0, t_1]$ such that $|\hat{f}(x(\cdot), u(t), t)| \leqq m(t)$, $|d\hat{f}[x(\cdot), u(t), t; \cdot]| \leqq m(t)$ for each $x \in C([t_0 - h, t_1], X)$ and $t \in [t_0, t_1]$, where $d\hat{f}$ is the Fréchet differential of $\hat{f}$ with respect to x (see [2, pp. 1–2]). Defining $x^0(t) \equiv \int_{t_0}^t f^0(x(s), u(s), s)\, ds$ and $\hat{x} = (x^0, x)$, we have the following necessary conditions that must be satisfied by solutions to the above problem.

THEOREM 5.1. *Let (x^*, u^*) be optimal. Then there exist $\alpha^0 \leqq 0$, $\hat{\psi}: [t_0, \infty) \to R^{n+1}$, $\lambda: R^1 \to R^n$ with $\hat{\psi}$, λ of bounded variation and left continuous such that*

(i) *$\lambda = (\lambda_{n+1} - \lambda_1, \lambda_{n+2} - \lambda_2, \cdots, \lambda_{2n} - \lambda_n)$, where the λ_j, $j = 1, \cdots, 2n$, satisfy: λ_j is constant on $(-\infty, -h]$, $\lambda_j(s) = 0$ for $s > 0$, λ_j is left continuous and nonincreasing with*

$$|\alpha^0| + \sum_{j=1}^{2n} \operatorname{Var}([-h, 0]; \lambda_j) > 0;$$

(ii) *$\hat{\psi} = (\psi^0, \psi)$ satisfies $\psi^0 = \alpha^0 \leqq 0$ and*

$$\psi(s) = \begin{cases} \displaystyle\int_{s-t_1}^{0^+} d\lambda(\theta) + \int_s^{t_1^+} d\psi(\theta)\mu(\theta, s) - \int_s^{t_1} [\psi(\theta)\eta^*(\theta, s) \\ \qquad - \alpha^0 f_x^{0*}(\theta)]\, d\theta, \quad s \in [t_0, t_1], \\ 0, \quad s > t_1, \end{cases}$$

where $f_x^{0*}(\theta) = f_x^0(x^*(\theta), u^*(\theta), \theta)$ and η^* is such that $df[x^*(\cdot), u^*(t), t; y]$ $= \int_{t_0-h}^t d_s\eta^*(t, s)y(s)$ for $t \in [t_0, t_1]$, $y \in \mathscr{C}^n[t_0 - h, t_1]$;

(iii) *$\int_{t_0}^{t_1} \hat{\psi}(s)\hat{f}(x^*(\cdot), u(s), s)\, ds \leqq \int_{t_0}^{t_1} \hat{\psi}(s)\hat{f}(x^*(\cdot), u^*(s), s)\, ds$ for all $u \in \mathscr{U}$.*

Proof. Let $\mathscr{E} = \mathscr{E}' \equiv \{\hat{x} \in \mathscr{C}^{n+1}[t_0 - h, t_1] | \dot{x}^0(t) = f^0(x(t), u(t), t), (d/dt)D(x(\cdot), t) = f(x(\cdot), u(t), t), t \in [t_0, t_1]$, for some $u \in \mathscr{U}$, and $\hat{x}_{t_0} = (0, \varphi)\}$, $\mathscr{Z}_0 \equiv \mathscr{C}^1[-h, 0]$, and $Z_0 = \{y \in \mathscr{Z}_0 | y(\theta) < 0, \theta \in [-h, 0]\}$. Define mappings $\varphi_0 : \mathscr{E} \to R^1$, $\varphi_{-i} : \mathscr{E} \to \mathscr{Z}_0$, $i = -1, -2, \cdots, -2n$, by $\varphi_0(\hat{x}) = x^0(t_1)$, $\varphi_{-i}(\hat{x})(\theta) = \zeta^i(\theta) - x^i(t_1 + \theta)$, $\varphi_{-i-n}(\hat{x})(\theta) = x^i(t_1 + \theta) - \zeta^i(\theta)$, $\theta \in [-h, 0]$, $i = 1, 2, \cdots, n$, where $\zeta = (\zeta^1, \cdots, \zeta^n)$. The problem formulated above is then equivalent to the problem of finding $\hat{z} \in \mathscr{E}$ such that $\varphi_i(\hat{z}) \in \overline{Z}_0$, $i = -1, -2, \cdots, -2n$, $\varphi_0(\hat{z}) \leqq \varphi_0(\hat{x})$ for all $\hat{x} \in \mathscr{E}$ with $\varphi_i(\hat{x}) \in \overline{Z}_0$, $i = -1, \cdots, -2n$. Letting $\mathscr{Z} \equiv R^1 \times \prod_{i=1}^{2n} \mathscr{Z}_0$, $Z \equiv \{\alpha \in R^1 | \alpha < 0\} \times \prod_{i=1}^{2n} Z_0$, and $\phi = (\varphi_0 - \varphi_0(\hat{z}), \varphi_{-1}, \cdots, \varphi_{-2n})$, one has that a solution to this problem is a (ϕ, Z) extremal [47, Definition 2.2]. Defining $\mathscr{E}_1 \equiv \mathscr{E}$ and $\mathscr{Y} \equiv \mathscr{C}^{n+1}[t_0 - h, t_1]$, it is obvious that Condition 6.1 of [47, p. 75] obtains. Next let $\hat{f}^*(t) = \hat{f}(x^*(\cdot), u^*(t), t)$, $\hat{\mathfrak{F}} \equiv \{\hat{g} : [t_0, t_1] \to R^{n+1} | \hat{g}(t) = \hat{f}(x^*(\cdot), u(t), t)$ for some $u \in \mathscr{U}\}$, and let $\hat{Y}$ be the solution to

$$\begin{aligned} &\hat{Y}(s, t) = E_{n+1} + \int_s^{t^+} d_\alpha \hat{Y}(\alpha, t)\hat{\mu}(\alpha, s) - \int_s^t \hat{Y}(\alpha, t)\hat{\eta}(\alpha, s)\, d\alpha, \quad s < t, \\ &\hat{Y}(t, t) = E_{n+1}, \qquad \hat{Y}(s, t) = 0, \qquad s > t, \end{aligned} \tag{5.1}$$

where E_{n+1} is the identity in $\mathscr{L}_{(n+1)\times(n+1)}$,

$$\hat{\mu}(\alpha, s) = \begin{pmatrix} 0 & 0 \\ 0 & \mu(\alpha, s) \end{pmatrix}, \qquad \hat{\eta}(\alpha, s) = \begin{pmatrix} 0 & -f_x^{0*}(\alpha) \\ 0 & \eta^*(\alpha, s) \end{pmatrix}.$$

Define the convex set $\mathscr{M}$ by $\mathscr{M} \equiv \{\widehat{\delta x} \in \mathscr{C}^{n+1}[t_0 - h, t_1] | \widehat{\delta x}(t) = 0, t \in [t_0 - h, t_0]$, $\widehat{\delta x}(t) = \int_{t_0}^t \hat{Y}(\alpha, t)\, \widehat{\delta f}(\alpha)\, d\alpha$, $t \in [t_0, t_1]$, $\widehat{\delta f} \in \operatorname{co}(\hat{\mathfrak{F}}) - \hat{f}^*\}$. Using the hypotheses of this paper, a generalized idea of quasiconvexity [1], [2], [21], [46], [47], the chattering lemma of Gamkrelidze [21, Lemma 4.1], [2, Lemma 3.1], and reasoning similar to that of Neustadt [46, § 3], a long and tedious argument verifies that Condition 6.2′ of [47, p. 76] is satisfied. For retarded systems the arguments are much like those in [46]; see also [1], [2]. Some technical difficulties are involved for neutral systems, but Kent [39] has shown that these can be overcome. Finally, Condition 6.4 of [47] can easily be shown to hold with $h = (h_0, h_{-1}, \cdots, h_{-2n})$ mapping $\mathscr{Y}$ into $\mathscr{Z}$ given by $h_0(\hat{y}) \equiv y^0(t_1)$, $h_{-i}(\hat{y})(\theta) \equiv -y^i(t_1 + \theta)$, $h_{-i-n}(\hat{y})(\theta) \equiv y^i(t_1 + \theta)$, $\theta \in [-h, 0]$, $i = 1, 2, \cdots, n$. Thus, Theorem 6.2 and hence Theorem 3.1 of [47] hold. We shall show that the desired results follow from Theorem 3.1.

Applying Theorem 3.1 of [47] we obtain the existence of a nonzero $l \in \mathscr{Z}^*$ such that

$$l(\hat{x}) \geqq 0 \quad \text{for all } \hat{x} \in \overline{Z}, \tag{5.2}$$

$$l \circ h(\widehat{\delta x}) \leqq 0 \quad \text{for all } \widehat{\delta x} \in \mathscr{M}. \tag{5.3}$$

From the definition of $\mathscr{Z}$, we see that there are $\alpha^0 \in R^1$ and $l_{-1} \in \mathscr{Z}_0^*$, $i = 1, 2, \cdots, 2n$, such that

$$l(\omega, \xi_1, \cdots, \xi_{2n}) = \alpha^0 \omega + \sum_{i=1}^{2n} l_{-i}(\xi_i)$$

for every $(\omega, \xi_1, \cdots, \xi_{2n}) \in \mathscr{Z}$. Since $\mathscr{Z}_0 = \mathscr{C}^1[-h, 0]$, it follows that there are $\lambda_i : (-\infty, \infty) \to R^1$ of bounded variation, with λ_i left continuous, $\lambda_i(s) = 0$ for

$s > 0$, λ_i constant on $(-\infty, -h]$, $i = 1, 2, \cdots, 2n$, such that

$$l_{-i}(\xi) = \int_{-h}^{0^+} d\lambda_i(\theta)\xi(\theta)$$

for every $\xi \in \mathscr{C}^1[-h, 0]$. Observing that $(-1, 0, \cdots, 0), (0, \xi, 0, \cdots, 0), \cdots, (0, 0, \cdots, \xi)$ are in $\overline{Z}$ whenever $\xi \in \overline{Z}_0$ we conclude from (5.2) that $\alpha^0 \leqq 0$ and λ_i is nonincreasing, $i = 1, 2, \cdots, 2n$. The fact that l is nonzero yields

$$|\alpha^0| + \sum_{i=1}^{2n} \operatorname{Var}([-h, 0]; \lambda_i) > 0.$$

The inequality (5.3) thus becomes

$$\alpha^0\, \delta x^0(t_1) + \sum_{i=1}^{n} \int_{-h}^{0^+} d\lambda_i(\theta)[-\delta x^i(t_1 + \theta)] + \sum_{i=1}^{n} \int_{-h}^{0^+} d\lambda_{n+i}(\theta)[\delta x^i(t_1 + \theta)] \leqq 0$$

for all $\widehat{\delta x} \in \mathscr{M}$. Defining $\lambda: R^1 \to R^n$ as in statement (i) above, this then becomes

$$\alpha^0\, \delta x^0(t_1) + \int_{-h}^{0^+} d\lambda(\theta)\, \delta x(t_1 + \theta) \leqq 0 \tag{5.4}$$

for all $\widehat{\delta x} \in \mathscr{M}$. The matrix solution $\hat{Y}$ in (5.1) may be written $\hat{Y} = \begin{pmatrix} 1 & Y^0 \\ 0 & Y \end{pmatrix}$, where $Y^0(s, t)$ is an n-dimensional row vector and $Y(s, t)$ is in $\mathscr{L}_{n\times n}$. The equations for Y^0, Y are

$$Y^0(s, t) = \int_s^{t^+} d_\alpha Y^0(\alpha, t)\mu(\alpha, s) - \int_s^t [-f_x^0{}^*(\alpha) + Y^0(\alpha, t)\eta^*(\alpha, s)]\, d\alpha$$

and

$$Y(s, t) = E_n + \int_s^{t^+} d_\alpha Y(\alpha, t)\mu(\alpha, s) - \int_s^t Y(\alpha, t)\eta^*(\alpha, s)\, d\alpha$$

for $s \leqq t$ with $Y^0(s, t)$, $Y(s, t)$ vanishing for $s > t$. Using the definition of $\mathscr{M}$ in (5.4) we obtain

$$\alpha^0 \int_{t_0}^{t_1} [\delta f^0(s) + Y^0(s, t_1)\, \delta f(s)]\, ds + \int_{-h}^{0^+} d\lambda(\theta)\left[\int_{t_0}^{t_1+\theta} Y(s, t_1 + \theta)\, \delta f(s)\, ds\right] \leqq 0$$

for all $\widehat{\delta f}$ in $\operatorname{co}(\tilde{\mathfrak{F}}) - \hat{f}^*$. This can be written as

$$\int_{t_0}^{t_1} \left[\alpha^0\, \delta f^0(s) + \left\{\alpha^0 Y^0(s, t_1) + \int_{-h}^{0^+} d\lambda(\theta) Y(s, t_1 + \theta)\right\} \delta f(s)\right] ds \leqq 0.$$

Defining $\psi(s) \equiv \alpha^0 Y^0(s, t_1) + \int_{-h}^{0^+} d\lambda(\theta) Y(s, t_1 + \theta)$ and $\psi^0 = \alpha^0$, an appropriate choice of $\widehat{\delta f}$ yields (iii). The properties of $\hat{Y}$ guaranteed by Theorem 3.1 of § 3 of this paper lead easily to the conclusion that ψ is left continuous and of bounded variation on $[t_0, \infty)$ with $\psi(s) = 0$ for $s > t_1$. Using the equations for Y^0, Y in the definition of ψ and making several interchanges in the orders of integration (it is

here that the Borel measurability of $\hat{Y}$ becomes important—see Remark 3.1 above) which are justified by the Fubini-type theorem in [10], we easily conclude that ψ satisfies the equation in (ii). This completes the proof of Theorem 5.1.

We point out that the equation for ψ in (ii) can be written in the equivalent form

$$\psi(s) = -\lambda(s - t_1) + \int_s^{t_1^+} d\psi(\theta)\mu(\theta, s) - \int_s^{t_1} [\psi(\theta)\eta^*(\theta, s) - \alpha^0 f_x^{0*}(\theta)]\, d\theta.$$

Since λ is usually only of bounded variation (BV) on $[-h, 0]$, one would scarcely expect ψ to be smoother, say absolutely continuous (AC), on $[t_1 - h, t_1]$. But for $s < t_1 - h$, the $\lambda(s - t_1)$ term is a constant ($\lambda(-h)$) and one might ask whether the equation can be written in differentiated form (i.e., is ψ AC?) for $s \in [t_0, t_1 - h]$. Even for *simple* neutral systems ψ is not in general AC. For example, if the system in the problem has the form (2.3), the equation for ψ becomes

$$\psi(s) = -\lambda(s - t_1) + \psi(s + h)A(s + h) + \int_{s+h}^{t_1} \psi(\theta)C(\theta)\, d\theta$$

$$+ \int_s^{t_1} [\psi(\theta)B(\theta) + \alpha^0 f_x^{0*}(\theta)]\, d\theta,$$

and it is easily seen that ψ has jump discontinuities at $s = t_1 - h, t_1 - 2h, t_1 - 3h, \cdots$. However, for problems involving system (2.4) the ψ equation is

$$\psi(s) = -\lambda(s - t_1) + \int_{s+h}^{t_1} \psi(\theta)C(\theta)\, d\theta + \int_s^{t_1} [\psi(\theta)B(\theta) + \alpha^0 f_x^{0*}(\theta)]\, d\theta,$$

and ψ is readily seen to be AC on $[t_0, t_1 - h]$ satisfying a.e.

$$\dot{\psi}(s) = -\alpha^0 f_x^{0*}(s) - \psi(s)B(s) - \psi(s + h)C(s + h).$$

In fact, there are a number of types of retarded systems for which the corresponding ψ will be AC on $[t_0, t_1 - h]$. These are discussed in [2, pp. 15–16]. Note that for retarded systems the associated ψ equation is the same as the $\bar{\lambda}$ equation of (i) of Theorem 1 in [2, p. 3] with the exception of the $\lambda(s - t_1)$ term which will be constant for $s < t_1 - h$.

If, instead of $\zeta \in \mathscr{C}^n[-h, 0]$, one only specifies q components ($q < n$) of x on $[t_1 - h, t_1]$, then one can derive the corresponding necessary conditions as above by using $2q$ of the φ_{-i} constraint functions instead of $2n$. The ideas for use of the "conflict" constraints φ_{-i} evolved from trying to treat the terminal conditions in $\mathscr{C}^n[-h, 0]$ as some type of bounded state variable restraint. It is not surprising then that in a number of examples [39], the multiplier λ behaves much like the multipliers obtained in considering bounded state variable problems [8], [48]. That is, λ has jumps at $t_1 - h, t_1$ (the points at which the trajectory enters and leaves the "boundary" ζ of the restraint region), and has jumps at points where the trajectory follows the "boundary" across a point where it is not smooth.

The multiplier λ can also be interpreted as a terminal boundary condition in function space [36], [38] which usually results from transversality conditions.

Let us mention briefly other ways of deriving the necessary conditions of Theorem 5.1. First, various other types of "conflicting" constraints [39] will yield essentially the same theorem; also it is possible to use a constraint of the form $\varphi_{-1}(\hat{x}) \equiv |x_{t_1} - \zeta| - \varepsilon$ (an ε-ball about ζ in $\mathscr{C}^n[-h, 0]$) together with Theorem 5.1 of [47]. Allowing $\varepsilon \to 0$ and using the linear "subfunctionals" Λ of that theorem yield (formally) the same necessary conditions as derived above. All of these proofs yield necessary conditions that have a notable deficiency. Observe that for the maximum principle (iii) to be nontrivial, one needs α^0 or λ or both nonzero. The nonzero statement involving l in the proof (see also (i) of the theorem) does not guarantee this. In fact, nothing rules out the situation $\alpha^0 = 0$ and $\lambda_i = \lambda_{n+i}$, $i = 1, 2, \cdots, n$, in which case the inequality in (iii) is trivially true with $\psi \equiv 0$, $\alpha^0 = 0$. This difficulty also appears in the approach employing Theorem 5.1 of [47] mentioned above (there is no guarantee that $\Lambda \neq 0$). Nonetheless, as we shall see, the conditions of Theorem 5.1 above are nontrivial in numerous examples [39] and are actually sufficient for linear problems with certain payoffs when $\alpha^0 \neq 0$ (normality). Thus we do obtain conditions which are both necessary and sufficient for a nonempty class of normal problems.

There are derivations of the necessary conditions presented here which for special problems do yield $\alpha^0 \neq 0$ (and hence nontriviality and sufficiency). The authors have shown that using an attainable sets approach [42] in function space for linear retarded systems with integral quadratic payoff and $\mathscr{U} = L_2$, necessary conditions which are the same as those of Theorem 5.1 can be obtained. Jacobs and Kao [36], [38] have obtained equivalent necessary conditions for problems with general nonlinear retarded systems by employing an abstract Lagrange multiplier rule [43, p. 243] in the function space $W_2^{(1)}$. But both of these latter approaches are for unconstrained controls and require more restrictive hypotheses than are desirable (roughly speaking, the matrix $D(t)$, where $k(u(t), t) = D(t)u(t)$, has rank n for a.e. t).

We next exhibit a class of problems for which the necessary conditions derived above are also sufficient. Let $f^0(z, u, t) = g^0(z, t) + k^0(u, t)$ and $f(x(\cdot), u, t) = \int_{t_0-h}^{t} d_s\eta(t, s)x(s) + k(u, t)$ satisfy the hypotheses given preceding Theorem 5.1, where η is as in Assumption 3.3. In addition, assume that $g^0: R^n \times [t_0, t_1] \to R^1$ is convex in z for each fixed $t \in [t_0, t_1]$. Under these hypotheses we have the following sufficiency results.

THEOREM 5.2. *Suppose* (x^*, u^*) *satisfy the conditions of Theorem* 5.1 *with* $\psi^0 = \alpha^0 < 0$. *Then* (x^*, u^*) *are optimal.*

Proof. Let $v \in \mathscr{U}$ be such that the corresponding trajectory $x = x(v)$ satisfies $x_{t_1} = \zeta = x^*_{t_1}$. Then from the equations for x, x^* we have

$$0 = \int_{t_0}^{t_1^+} d\psi(s)\Big\{[x(s) - x^*(s)] - \int_{t_0}^{t_1^+} d_\theta\mu(s, \theta)[x(\theta) - x^*(\theta)]$$

$$- \int_{t_0}^{s}\int_{t_0}^{t_1^+} d_\theta\eta(\sigma, \theta)[x(\theta) - x^*(\theta)]\,d\sigma - \int_{t_0}^{s}[k(v(\sigma), \sigma) - k(u^*(\sigma), \sigma)]\,d\sigma\Big\}.$$

Then, using the definition of x^0, we obtain

$$
\begin{aligned}
-\alpha^0[x^{0*}(t_1) - x^0(t_1)] = \alpha^0 &\int_{t_0}^{t_1} [g^0(x(s), s) - g^0(x^*(s), s)]\, ds \\
&+ \alpha^0 \int_{t_0}^{t_1} [k^0(v(s), s) - k^0(u^*(s), s)]\, ds \\
&+ \int_{t_0}^{t_1^+} d\psi(s)[x(s) - x^*(s)] \\
&- \int_{t_0}^{t_1^+} d\psi(s) \int_{t_0}^{t_1^+} d_\theta \mu(s, \theta)[x(\theta) - x^*(\theta)] \\
&- \int_{t_0}^{t_1^+} d\psi(s) \left\{ \int_{t_0}^{s} \int_{t_0}^{t_1^+} d_\theta \eta(\sigma, \theta)[x(\theta) - x^*(\theta)]\, d\sigma \right\} \\
&- \int_{t_0}^{t_1^+} d\psi(s) \int_{t_0}^{s} [k(v(\sigma), \sigma) - k(u^*(\sigma), \sigma)]\, d\sigma.
\end{aligned}
$$

Adding and subtracting a term involving $g_x^{0*}(s) \equiv g_x^0(x^*(s), s)$, integrating by parts in the last two terms, and noting that $\psi(t_1^+) = 0$, we find

$$
\begin{aligned}
-\alpha^0[x^{0*}(t_1) - x^0(t_1)] = \alpha^0 &\int_{t_0}^{t_1} \{g^0(x(s), s) - g^0(x^*(s), s) - g_x^{0*}(s)[x(s) - x^*(s)]\}\, ds \\
&+ \int_{t_0}^{t_1} \left\{ \alpha^0[k^0(v(s), s) - k^0(u^*(s), s)] + \psi(s)[k(v(s), s) - k(u^*(s), s)] \right\} ds \\
&+ \int_{t_0}^{t_1^+} d\psi(s)[x(s) - x^*(s)] - \int_{t_0}^{t_1^+} d\psi(s) \int_{t_0}^{t_1^+} d_\theta \mu(s, \theta)[x(\theta) - x^*(\theta)] \\
&+ \int_{t_0}^{t_1} \psi(s) \int_{t_0}^{t_1^+} d_\theta \eta(s, \theta)[x(\theta) - x^*(\theta)]\, ds + \alpha^0 \int_{t_0}^{t_1} g_x^{0*}(s)[x(s) - x^*(s)]\, ds.
\end{aligned}
$$

Using the Fubini-type theorem [10] to interchange the order of integration in several of the integrals and combining terms, we have

$$
\begin{aligned}
-\alpha^0[x^{0*}(t_1) - x^0(t_1)] = \alpha^0 &\int_{t_0}^{t_1} \{g^0(x(s), s) - g^0(x^*(s), s) - g_x^{0*}(s)[x(s) - x^*(s)]\}\, ds \\
&+ \int_{t_0}^{t_1} \hat{\psi}(s)[\hat{f}(x^*(\cdot), v(s), s) - \hat{f}(x^*(\cdot), u^*(s), s)]\, ds \\
&+ \int_{t_0}^{t_1^+} d_\theta \left\{ \psi(\theta) - \int_{t_0}^{t_1^+} d\psi(s)\mu(s, \theta) + \int_{t_0}^{t_1} \psi(s)\eta(s, \theta)\, ds \right. \\
&\qquad \left. - \int_{\theta}^{t_1} \alpha^0 g_x^{0*}(s)\, ds \right\} [x(\theta) - x^*(\theta)].
\end{aligned}
$$

Noting that

$$
\int_{t_0}^{t_1^+} d\psi(s)\mu(s, \theta) = \int_{\theta}^{t_1^+} d\psi(s)\mu(s, \theta) \quad \text{and} \quad \int_{t_0}^{t_1} \psi(s)\eta(s, \theta)\, ds = \int_{\theta}^{t_1} \psi(s)\eta(s, \alpha)\, ds,
$$

it follows from the convexity assumption on g^0 and parts (ii) and (iii) of Theorem 5.1 that

$$\begin{aligned} -\alpha^0[x^{0*}(t_1) - x^0(t_1)] &\leqq \int_{t_0}^{t_1^+} d_\theta\lambda(\theta - t_1)[x(\theta) - x^*(\theta)] \\ &= \int_{t_1-h}^{t_1^+} d_\theta\lambda(\theta - t_1)[x(\theta) - x^*(\theta)] \\ &= 0, \end{aligned}$$

or since $\alpha^0 < 0$, $[x^{0*}(t_1) - x^0(t_1)] \leqq 0$. Thus (x^*, u^*) is optimal.

The underlying idea for this sufficiency proof can be traced at least as far back as a paper by Rozonoer [50]. A number of other authors [20], [22], [41], [42] have used and refined it. In fact, in a recent work of Funk and Gilbert [20] it is pointed out that under normality and certain convexity assumptions, the abstract necessary conditions derived by Neustadt [46], [47] are also sufficient in a number of situations.

6. Examples. We first present two examples which illustrate use of some of the results obtained above.

Example 6.1. Consider a system described by the scalar equation

$$\dot{x}(t) = x(t) - x(t-1) + u(t).$$

We wish to drive from the initial function $\varphi = 0$ to the target function given by $\zeta(\theta) = 2 + \theta, \theta \in [-1, 0]$, while minimizing $J(u) = \int_0^3 x(t)\,dt$ with $U = [-1, 1]$.

Since $f^0 = x$, the maximum condition of Theorem 5.1 reduces to

$$\int_0^3 \psi(s)u(s)\,ds \leqq \int_0^3 \psi(s)u^*(s)\,ds.$$

Thus $u^*(s) = \operatorname{sgn}[\psi(s)]$ whenever $\psi(s) \neq 0$. The equation for ψ reduces to

$$\psi(s) = -\lambda(s-3) + \int_{s+1}^3 \psi(\theta)(-1)\,d\theta + \int_s^3 [\psi(\theta)(1) + \alpha^0]\,d\theta, \qquad s \in [0, 3].$$

From the discussion and examples of § 4, we guess that u^* is not bang-bang on $[2, 3]$. Thus we try $\psi \equiv 0$ on $(2, 3]$ and $\alpha^0 = -1$. On $(2, 3]$ the equation for ψ becomes

$$0 = -\lambda(s-3) + \int_s^3 (-1)\,d\theta,$$

so

$$\lambda(\theta) = \theta \qquad \text{for } \theta \in (-1, 0].$$

On $[1, 2]$ we differentiate the equation for ψ and obtain

$$\dot{\psi}(s) = -\psi(s) - \alpha^0 = -\psi(s) + 1.$$

Thus $\psi(s) = c\,e^{-(s-2)} + 1$ on $[1, 2]$. But $\psi(2) = -\lambda(-1) + \int_2^3 (-1)\,d\theta = -\lambda(-1) - 1$, so $c + 1 = -1 - \lambda(-1)$. Let $v = -\lambda(-1)$; then $c = v - 2$, and

$$\psi(s) = (v - 2)\,e^{(2-s)} + 1, \qquad s \in [1, 2].$$

Similarly, on $[0, 1]$, $\dot{\psi}(\theta) = -\psi(\theta) + \psi(\theta + 1) + 1$. Integrating from $\psi(1) = (v - 2)e + 1$, we obtain

$$\psi(\theta) = (v - 2)\{e^{(2-\theta)} - (1 - \theta)\,e^{(1-\theta)}\} + 2 - e^{(1-\theta)}, \qquad \theta \in [0, 1].$$

We claim $\psi(0) < 0$. Suppose $\psi(0) \geqq 0$. Then $(v - 2)[e^2 - e] + 2 - e \geqq 0$, or $(v - 2) \geqq (e - 2)/(e^2 - e) > 0$. Thus $\psi(s) = (v - 2)\,e^{(2-s)} + 1 > 1$ for $s \in [1, 2]$, $\dot{\psi}(\theta) > -\psi(\theta) + 1 + 1 = -\psi(\theta) + 2$ for $\theta \in [0, 1]$. Hence if $\psi(0) \geqq 0$, then $\psi(s) > 0$ on $(0, 2)$, and $u^*(s) = +1$ for $s \in (0, 2)$. A simple integration shows that this implies $x^*(2) > 1$, contradicting the boundary condition $x(2) = 1$ and proving the claim. We next claim that $\psi(1) > 0$. Suppose $\psi(1) \leqq 0$. Then $(v - 2)e + 1 \leqq 0$, or $v - 2 \leqq -1/e$. For $\theta \in (0, 1)$,

$$\begin{aligned}\psi(\theta) &= (v - 2)[e^{(1-\theta)}\{e + \theta - 1\}] + 2 - e^{(1-\theta)} \\ &\leqq -e^{-1}[e^{(1-\theta)}\{e + \theta - 1\}] + 2 - e^{(1-\theta)} \\ &= -e^{-\theta}\{e + \theta - 1 + e\} + 2 = -e^{-\theta}\{2e + \theta - 1\} + 2.\end{aligned}$$

This expression has its maximum over $[0, 1]$ at $\theta = 1$,

$$\psi(1) \leqq -e^{-1}\{2\,e^1\} + 2 = 0,$$

and hence $\psi(\theta) < 0$ for $\theta \in (0, 1)$. For $s \in (1, 2)$,

$$\begin{aligned}\dot{\psi}(s) &= -\psi(s) + 1 = -(v - 2)\,e^{(2-s)} \\ &\geqq e^{-1}\,e^{(2-s)} = e^{(1-s)} > 0.\end{aligned}$$

Thus ψ has at most one zero in $[1, 2]$. If there is no zero in $[1, 2]$, $u^*(s) = -1$ for all $s \in (0, 2)$, and a simple integration shows that the response to such a control does not satisfy $x^*(2) = 1$. If ψ has a zero at $\omega \in [1, 2]$, then u^* is given on $(0, 2)$ by

$$u^*(s) = \begin{cases} -1, & s \in (0, \omega), \\ +1, & s \in (\omega, 2). \end{cases}$$

Another integration shows that the response to such a control also cannot satisfy $x^*(2) = 1$. Thus $\psi(1) > 0$ must hold.

From $\psi(1) > 0$ it follows that $(v - 2)e^1 + 1 > 0$, or $(v - 2)\,e^1 > -1$. On $(1, 2)$,

$$\begin{aligned}\psi(s) &= (v - 2)\,e^{(2-s)} + 1 = (v - 2)\,e^1\,e^{(1-s)} + 1 \\ &> -e^{(1-s)} + 1,\end{aligned}$$

so $\psi(s) > 0$ for $s \in (1, 2)$. On $(0, 1)$,

$$\dot{\psi}(\theta) = -\psi(\theta) + \psi(\theta + 1) + 1 > -\psi(\theta) + 1.$$

Thus we have that ψ has at most one zero in $(0, 1)$. Since ψ is continuous with $\psi(0) < 0$ and $\psi(1) > 0$, ψ has a zero in $(0, 1)$, which we denote by ω. This implies

the optimal control is given by

$$u^*(s) = \begin{cases} -1, & s \in (0, \omega), \\ +1, & s \in (\omega, 2). \end{cases}$$

The response to this control is

$$x^*(t) = \begin{cases} 1 - e^t, & t \in [0, \omega], \\ 2\,e^{(t-\omega)} - e^t - 1, & t \in [\omega, 1], \\ 2\,e^{(t-\omega)} - e^t + (t-2)\,e^{(t-1)}, & t \in [1, 1+\omega], \\ 2\,e^{(t-\omega)} - e^t + (t-2)\,e^{(t-1)} - 2(t-2-\omega)\,e^{(t-1-\omega)} - 2, & \\ \qquad\qquad t \in [1+\omega, 2]. \end{cases}$$

From the boundary conditions,

$$\begin{aligned} 1 - x^*(2) &= 2\,e^{(2-\omega)} - e^2 + 2\omega\,e^{(1-\omega)} - 2, \\ 2(e+\omega)\,e^{(1-\omega)} &= 3 + e^2, \\ (e+\omega)\,e^{(1-\omega)} &= \tfrac{1}{2}(3+e^2). \end{aligned}$$

The solution is approximately $\omega = 0.531$. To determine v we go back to the equation $\psi(\omega) = 0$:

$$\begin{aligned} &(v-2)\{e^{(2-\omega)} - (1-\omega)\,e^{(1-\omega)}\} + 2 - e^{(1-\omega)} = 0, \\ &v = 2 - \frac{2 - e^{(1-\omega)}}{e^{(2-\omega)} - (1-\omega)\,e^{(1-\omega)}} \approx 2 - 0.111 = 1.889. \end{aligned}$$

We determine u^* on (2, 3) by using the equation for x and the fact that $x(s) = s - 1$, $s \in (2, 3)$. On (2, 3), $1 = s - 1 - x(s-1) + u^*(s)$, $u^*(s) = 2 - s + x(s-1)$, from which it follows that

$$u^*(s) = \begin{cases} 2 - s + 2\,e^{(s-1-\omega)} - e^{(s-1)} + (s-3)\,e^{(s-2)}, & s \in (2, 2+\omega), \\ -s + 2\,e^{(s-1-\omega)} - e^{(s-1)} + (s-3)\,e^{(s-2)} - 2(s-3-\omega)\,e^{(s-2-\omega)}, & \\ \qquad\qquad s \in (2+\omega, 3). \end{cases}$$

Since $|u^*| \leqq 1$ this is an admissible control satisfying the necessary conditions with $\alpha^0 \neq 0$, and by Theorem 5.2, u^* is an optimal control. It is not difficult to argue that u^* is in fact the unique optimal control.

Example 6.2. Consider a system described by the scalar equation

$$\dot{x}(t) = \dot{x}(t-1) + u(t).$$

We wish to drive from the initial function $\varphi = 0$ to the target function given by

$$\zeta(\theta) = \begin{cases} -\frac{1}{2} - \theta, & \theta \in [-1, -\frac{1}{2}], \\ \frac{1}{2} + \theta, & \theta \in [-\frac{1}{2}, 0], \end{cases}$$

while minimizing $J(u) = \int_0^3 u^2(t)\,dt$ with $U = R^1$.

The maximum condition of Theorem 5.1 reduces to

$$\alpha^0[v^2 - u^{*2}(s)] + \psi(s)[v - u^*(s)] \leqq 0$$

for all $v \in R^1$, almost all $s \in [0, 3]$. Let us choose $\alpha^0 = -1$. Then this implies

$$u^*(s) = \tfrac{1}{2}\psi(s).$$

The equation for ψ reduces to

$$\psi(s) = \psi(s + 1) - \lambda(s - 3), \qquad s \in [0, 3];$$

hence we have that

$$u^*(s) = u^*(s + 1) - \frac{\lambda(-1)}{2} \quad \text{for } s \in (0, 2).$$

From the shape of the target, the equations, and this relationship between $u^*(s)$ and $u^*(s + 1)$, we guess that the optimal trajectory is composed of straight-line segments of time-length $\frac{1}{2}$. Let the slopes of these be $\alpha, \beta, \gamma, \delta$ respectively on $[0, 2]$, and set $-\frac{1}{2}\lambda(-1) = \kappa$. Thus

$$u^*(s) = \begin{cases} \alpha, & s \in (0, \frac{1}{2}), \\ \beta, & s \in (\frac{1}{2}, 1), \\ \gamma - \alpha, & s \in (1, \frac{3}{2}), \\ \delta - \beta, & s \in (\frac{3}{2}, 2), \\ -1 - \gamma, & s \in (2, \frac{5}{2}), \\ 1 - \delta, & s \in (\frac{5}{2}, 3). \end{cases}$$

Substituting these in $u^*(s) = u^*(s + 1) + \kappa$, we obtain the systems of equations

$$\delta - \beta = 1 - \delta + \kappa, \qquad \gamma - \alpha = -1 - \gamma + \kappa,$$
$$\beta = \delta - \beta + \kappa, \qquad \alpha = \gamma - \alpha + \kappa.$$

We solve in pairs: $\beta = 2\delta - 1 - \kappa$ and $\delta = 2\beta - \kappa$ imply $\beta = 4\beta - 2\kappa - 1 - \kappa$, or $3\beta = 3\kappa + 1$. Thus $\beta = \kappa + \frac{1}{3}$, $\delta = \kappa + \frac{2}{3}$. Also $\alpha = 2\gamma - \kappa + 1$ and $\gamma = 2\alpha - \kappa$ imply $\alpha = 4\alpha - 2\kappa + 1 - \kappa$, or $3\alpha = 3\kappa - 1$. Thus $\alpha = \kappa - \frac{1}{3}$, $\gamma = \kappa - \frac{2}{3}$. The endpoint $x^*(2) = \frac{1}{2}$ is half the sum of the slopes, so $\frac{1}{2} = \frac{1}{2}(\alpha + \beta + \gamma + \delta)$, $1 = 4\kappa$, or $\kappa = \frac{1}{4}$. Thus $\alpha = -\frac{1}{12}$, $\beta = \frac{7}{12}$, $\gamma = -\frac{5}{12}$, and $\delta = \frac{11}{12}$. From $u^*(s) = \frac{1}{2}\psi(s)$ and the expression for u^* in terms of α, β, γ, and δ we obtain

$$\psi(s) = \begin{cases} -\frac{1}{6}, & s \in (0, \frac{1}{2}), \\ \frac{7}{6}, & s \in (\frac{1}{2}, 1), \\ -\frac{4}{6}, & s \in (1, \frac{3}{2}), \\ \frac{4}{6}, & s \in (\frac{3}{2}, 2), \\ -\frac{7}{6}, & s \in (2, \frac{5}{2}), \\ \frac{1}{6}, & s \in (\frac{5}{2}, 3). \end{cases}$$

This ψ satisfies the equation in Theorem 5.1 with

$$\lambda(\theta) = \begin{cases} -\frac{1}{4}, & \theta \in (-\infty, -1], \\ \frac{7}{6}, & \theta \in (-1, -\frac{1}{2}], \\ -\frac{1}{6}, & \theta \in (-\frac{1}{2}, 0], \\ 0, & \theta \in (0, \infty). \end{cases}$$

The x^*, u^* obtained here satisfy the necessary conditions with ψ and λ given above and $\alpha^0 = -1$. By Theorem 5.2, u^* is optimal. Since the system equations are linear and f^0 is strictly convex in u, standard arguments [42] show that u^* is unique.

There are several ways in which one could attempt to solve fixed function target problems without using the necessary conditions for driving to a function. The most direct is to minimize $\int_{t_0}^{t_1-h} f^0(x(s), u(s), s)\, ds$ while driving to $x^*(t_1 - h)$, and then attempt to determine u^* on $(t_1 - h, t_1)$ so that $x^*_{t_1}$ is as desired. If $f^0(x, u, s)$ is independent of u for $s \in (t_1 - h, t_1)$ and $U = R^m$, this approach can succeed provided that given x on $[t_0 - h, t_1]$ one can solve the system equations for $u(s), s \in (t_1 - h, t_1)$. If $U \neq R^m$ the method may fail because the u^* determined on $(t_1 - h, t_1)$ in this manner need not lie in U; in the case of Example 6.1 it succeeds. In fact, given the assumption that $\psi(s) = 0$ for $s \in (t_1 - h, t_1]$ in Example 6.1 the subsequent arguments are identical to those resulting from applying the maximum principle for point-target problems [2], [40] as suggested above. If $f^0(x, u, s)$ is not independent of u for $s \in (t_1 - h, t_1)$, this approach will generally fail. In Example 6.2, driving to $x(2) = \frac{1}{2}$ while minimizing $\int_0^2 u^2(t)\, dt$ yields

$$u^*(s) = \begin{cases} \frac{1}{5}, & s \in (0, 1), \\ \frac{1}{10}, & s \in (1, 2), \end{cases} \qquad x^*(t) = \begin{cases} \frac{t}{5}, & t \in [0, 1], \\ \frac{3t - 1}{10}, & t \in [1, 2], \end{cases}$$

which is not even close to the optimal trajectory for the function-target problem.

A more complicated method involves making the function-target problem equivalent to a point-target problem by altering the form of f^0 on $[t_0, t_1 - h]$ to include the "cost" due to the u^* determined on $(t_1 - h, t_1)$ as above. With $U = R^m$ one may then use either the necessary conditions of the calculus of variations (see [16], [34], [51]) or the point-target maximum principle. Examples illustrating this can be found in [39]. In Example 6.2, the altered cost functional is

$$\begin{aligned} J(u) &= \int_0^1 u^2(t)\, dt + \int_1^{3/2} \{u^2(t) + [-1 - u(t) - u(t-1)]^2\}\, dt \\ &\quad + \int_{3/2}^2 \{u^2(t) + [1 - u(t) - u(t-1)]^2\}\, dt \\ &= \int_0^1 \dot{x}^2(t)\, dt + \int_1^{3/2} \{[\dot{x}(t) - \dot{x}(t-1)]^2 + [-1 - \dot{x}(t)]^2\}\, dt \\ &\quad + \int_{3/2}^2 \{[\dot{x}(t) - \dot{x}(t-1)]^2 + [1 - \dot{x}(t)]^2\}\, dt. \end{aligned}$$

With this form of $J(u)$ in terms of u, one clearly would prefer not to use the maximum principle on this equivalent problem. Applying the modified Euler conditions to the second expression for $J(u)$, one obtains that the optimal trajectory must consist of line segments of time-length 1/2. However, the relationships between the four slopes and a single constant are not obtained, and the problem of determining the slopes remains difficult without the function-target necessary conditions. When $U \neq R^m$ the problems encountered in the first point-target method will also occur here.

Acknowledgments. The authors would like to express appreciation to Professors J. K. Hale and M. Q. Jacobs for a number of stimulating discussions during the course of the work reported here.

Note added in proof. The authors have recently discovered that Kirillova, Curakova, and Gabasov have published a number of results on controllability of delayed systems to the zero function. These results [R. GABASOV AND S. V. CURAKOVA, Izv. Akad. Nauk SSSR Tehn. Kibernet., 4 (1969), pp. 17–28; R. GABASOV AND F. KIRILLOVA, *Qualitative Theory of Optimal Processes*, Moscow, 1971; F. M. KIRILLOVA AND S. V. CURAKOVA, *Differencial'nye Uravnenija*, 3 (1967), pp. 436–445; F. M. KIRILLOVA AND S. V. CURAKOVA, Dokl. Akad. Nauk SSSR, 174 (1967), pp. 1260–1263] have been discussed in a survey paper by one of the authors of this manuscript [H. T. BANKS, *Control of functional differential equations with function space boundary conditions*, Proc. Park City Differential Equations Symposium (March 7–11, 1972, Park City, Utah), Academic Press, New York, 1972].

REFERENCES

[1] H. T. BANKS, *Necessary conditions for control problems with variable time lags*, this Journal, 6 (1968), pp. 9–47.

[2] ———, *Variational problems involving functional differential equations*, this Journal, 7 (1969), pp. 1–17.

[3] ———, *Representations for solutions of linear functional differential equations*, J. Differential Equations, 5 (1969), pp. 399–409.

[4] H. T. BANKS AND M. Q. JACOBS, *The optimization of trajectories of linear functional differential equations*, this Journal, 8 (1970), pp. 461–488.

[5] H. T. BANKS, M. Q. JACOBS AND M. R. LATINA, *The synthesis of optimal controls for linear problems with retarded controls*, J. Optimization Theory Appl., 8 (1971), pp. 319–366.

[6] R. BELLMAN AND K. L. COOKE, *Differential Difference Equations*, Academic Press, New York, 1963.

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

[8] L. D. BERKOVITZ, *On control problems with bounded state variables*, J. Math. Anal. Appl., 5 (1962), pp. 488–498.

[9] R. K. BRAYTON, *Nonlinear oscillations in a distributed network*, Quart. Appl. Math., 24 (1966–67), pp. 283–301.

[10] R. H. CAMERON AND W. T. MARTIN, *An unsymmetric Fubini theorem*, Bull. Amer. Math. Soc., 47 (1941), pp. 121–125.

[11] R. CONTI, *A convex programming problem in Banach spaces and applications to optimum control theory*, J. Comput. System Sci., 4 (1970), pp. 38–49.

[12] K. L. COOKE AND D. W. KRUMME, *Differential-difference equations and nonlinear initial-boundary value problems for linear hyperbolic partial differential equations*, J. Math. Anal. Appl., 24 (1968), pp. 372–387.

[13] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. I, Interscience, New York, 1953.

[14] J. W. DETTMAN, *Mathematical Methods in Physics and Engineering*, McGraw-Hill, New York, 1962.

[15] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1958.

[16] L. E. EL'SGOL'TS, *Qualitative methods in mathematical analysis*, Translations of Mathematical Monographs 12, American Mathematical Society, Providence, 1964.

[17] P. L. FALB, *Infinite dimensional control problems. I: on the closure of the set of attainable states for linear systems*, J. Math. Anal. Appl., 9 (1964), pp. 12–22.

[18] H. O. FATTORINI, *A remark on the "bang-bang" principle for linear control systems in infinite-dimensional space*, this Journal, 6 (1968), pp. 109–113.

[19] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[20] J. E. FUNK AND E. G. GILBERT, *Some sufficient conditions for optimality in control problems with state space constraints*, this Journal, 8 (1970), pp. 498–504.

[21] R. V. GAMKRELIDZE, *On some extremal problems in the theory of differential equations with applications to the theory of optimal control*, this Journal, 3 (1965), pp. 106–128.

[22] A. HALANAY, *Le "principe du maximum" pour les systemes optimaux lineaires à retardement*, C. R. Acad. Sci. Paris, 254 (1962), pp. 2277–2279.

[23] J. K. HALE, *Linear functional-differential equations with constant coefficients*, Contributions to Differential Equations, 2 (1963), pp. 291–317.

[24] ———, *Functional Differential Equations*, Applied Mathematical Sciences, vol. 3, Springer-Verlag, New York, 1971.

[25] J. K. HALE AND M. A. CRUZ, *Existence, uniqueness, and continuous dependence for hereditary systems*, Ann. Mat. Pura Appl. (4), 85 (1970), pp. 63–81.

[26] J. K. HALE AND K. R. MEYER, *A class of functional equations of neutral type*, Mem. Amer. Math. Soc., 76 (1967), pp. 1–65.

[27] H. HALKIN, *A generalization of LaSalle's "bang-bang" principle*, this Journal, 2 (1964), pp. 199–202.

[28] ———, *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.

[29] 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.

[30] D. HENRY, *Linear FDE's of neutral type in the space of continuous functions: representation of solutions and the adjoint equation*, to appear.

[31] H. HERMES, *On the closure and convexity of attainable sets in finite and infinite dimensions*, this Journal, 5 (1967), pp. 409–417.

[32] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[33] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, American Mathematical Society, Providence, 1957.

[34] D. K. HUGHES, *Variational and optimal control problems with delayed argument*, J. Optimization Theory Appl., 2 (1968), pp. 1–14.

[35] M. Q. JACOBS, *Attainable sets in systems with unbounded controls*, J. Differential Equations, 4 (1968), pp. 408–423.

[36] M. Q. JACOBS AND T. J. KAO, *An optimum settling problem for time lag systems*, to appear.

[37] G. A. KAMENSKII AND E. A. KHVILON, *Necessary conditions for the optimal control of systems with deviating argument of neutral type*, Automat. Remote Control, 30 (1969), pp. 327–339.

[38] T. J. KAO, *Optimal control for functional differential equations*, Masters thesis, Brown University, Providence, 1971.

[39] G. A. KENT, *Optimal control of functional differential equations of neutral type*, Doctoral thesis, Brown University, Providence, 1971.

[40] ———, *A maximum principle for optimal control problems with neutral functional differential systems*, Bull. Amer. Math. Soc., 77 (1971), pp. 565–570.

[41] E. B. LEE, *A sufficient condition in the theory of optimal control*, this Journal, 1 (1963), pp. 241–245.

[42] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[43] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[44] A. MANITIUS, *Optimal control of processes with time delays—a survey and some new results*, Arch. Avtomat. i Telemekh., 15 (1970), pp. 205–221.
[45] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
[46] ———, *An abstract variational theory with applications to a broad class of optimization problems. II*, this Journal, 5 (1967), pp. 90–137.
[47] ———, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.
[48] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
[49] V. M. POPOV, *On the property of reachability for some delay-differential equations*, Tech. Res. Rep. R-70-08, Department of Electrical Engrg., University of Maryland, College Park, 1970.
[50] L. I. ROZONOER, *L. S. Pontryagin's Maximum Principle in the theory of optimum systems. I, II*, Automat. Remote Control, 20 (1959), pp. 1288–1302, 1405–1421.
[51] L. D. SABBAGH, *Variational problems with lags*, J. Optimization Theory Appl., 3 (1969), pp. 34–51.
[52] M. SLEMROD, *Nonexistence of oscillations in a nonlinear distributed network*, J. Math. Anal. Appl., 36 (1971), pp. 22–40.
[53] I. N. SNEDDON, *Elements of Partial Differential Equations*, McGraw-Hill, New York, 1957.
[54] L. WEISS, *On the controllability of delay-differential systems*, this Journal, 5 (1967), pp. 575–587.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

NECESSARY CONDITIONS FOR OPTIMAL CONTROLS OF ELLIPTIC OR PARABOLIC PROBLEMS*

TULLIO ZOLEZZI†

Abstract. In this paper I prove necessary conditions for optimal controls of boundary value parabolic or elliptic problems, when the coefficients of the operators depend on the controls. Some stochastic control problems can be formulated as optimal controls for the coefficients of elliptic or parabolic problems, and the necessary conditions proved here generalize known results to the case when the second order coefficients depend on the controls. From the necessary conditions uniqueness criteria and "bang-bang" theorems are deduced, and, suggested by the direct physical interpretation of these problems, stability (in a variational sense) theorems are proved.

Introduction. If the maximum order coefficients depend on the control, existence theorems for optimal controls of boundary value elliptic, or parabolic, problems are proved in [13]. In this paper I prove, in the general case, corresponding necessary conditions, that extend known results of [6] to the case when the second order coefficients of the differential operators depend on the control. (In terms of the corresponding stochastic perturbations of the Pontryagin problem in ordinary control theory, such a situation arises when one controls also the coefficient of the perturbation.) From such necessary conditions, uniqueness and (in particular cases) "bang-bang" theorems are deduced.

A direct physical interpretation of these control problems is the following: in the parabolic case, for example, we minimize the temperature distribution (or a more general functional of the temperature), in a given time, heating some materials, varying their thermophysical parameters. Therefore in the last section of the present paper theorems are proved, showing when such control problems are "well-posed" from a variational point of view.

1. Notations. R^m is the m-dimensional real space; "a.e." means almost everywhere with respect to Lebesgue measure; "meas." means Lebesgue measure. f is a Carathéodory function in $A \times B$ if $f(\cdot, b)$ is measurable for every $b \in B$ and $f(a, \cdot)$ is continuous a.e. in A. If $Q \equiv \Omega \times (0, T]$, where Ω is a bounded open set, whose boundary is $\partial\Omega$ and closure $\bar{\Omega}$, and if X is a Banach space of functions defined in Ω, $L^q(X)(1 \leqq q < \infty)$ denotes the space of functions f, measurable in Q, such that $f(\cdot, t) \in X$ a.e. in $(0, T]$ and $\int_0^T \|f(\cdot, t)\|_X^q \, dt < +\infty$: $L^\infty(X)$ is defined analogously. $H_0^1(\Omega)$ is the completion of $C_0^1(\Omega)$ (real continuously differentiable functions with compact support in Ω) in the norm $\|u\|_{L^2(\Omega)} + \sum_{k=1}^m \|u_{x_k}\|_{L^2(\Omega)}$. A° is the interior of A. $L^{p,q}(Q)$ denotes $L^q(L^p(\Omega))$. $\to$ ($\rightharpoonup$) denotes strong (weak) convergence in any Banach space. $\rightarrow$ denotes the L^1-convergence in L^∞ (weak-star convergence). $f(u)$ denotes the function $y \to f(y, u(y))$. $F(U)$ is $\{f(u) : u \in U\}$. f_x denotes the gradient of f in the variables $x \equiv (x_1, \cdots, x_n)$. "$\equiv$" means equal by definition. $d(y, A)$ is the distance of y from the set A. Any subsequence is denoted as the original sequence. "$f \neq 0$" means f is different from the function 0.

* Received by the editors September 22, 1970, and in revised form September 17, 1971.

† Istituto di Matematica, v. L. B. Alberti 4, 16132 Genova, Italy. This research was supported by Centro di Matematica e fisica teorica del C.N.R. presso l'Università di Genova.

2. Definitions and common assumptions. In the following Ω denotes a fixed open bounded set in R^m, $T > 0$, $Q \equiv \Omega \times (0, T]$.

If

$$L \equiv \frac{\partial}{\partial t} - \sum_{i,k=1}^{m} \frac{\partial}{\partial x_k}\left[a_{ik}(x,t)\frac{\partial}{\partial x_i} + a_k(x,t)\right] - \sum_{k=1}^{m} b_k(x,t)\frac{\partial}{\partial x_k} - c(x,t)$$

is a given parabolic operator in Q, recall that z is the (weak) solution, in the sense of [1], of the problem

$$\text{(I)} \qquad \begin{aligned} Lz &= f(x,t) \quad \text{in } Q, \\ z(x,0) &= z_0(x) \quad \text{in } \Omega, \qquad z(x,t) = 0 \quad \text{in } \partial\Omega \times [0,T], \end{aligned}$$

if $z \in L^\infty(L^2(\Omega)) \cap L^2(H_0^1(\Omega))$ and

$$\int_Q \left(-zh_t + \sum_{i,k=1}^{m} a_{ik} z_{x_i} h_{x_k} + \sum_{k=1}^{m} a_k z h_{x_k} - \sum_{k=1}^{m} b_k z_{x_k} h - czh - fh \right) dx\, dt = 0$$

for every $h \in C_0^1(Q^\circ)$, and if

$$\lim_{t \to 0+} \int_\Omega z(x,t)k(x)\, dx = \int_\Omega z_0 k\, dx$$

for every $k \in C_0^1(\Omega)$.

If

$$L \equiv -\sum_{i,k=1}^{m} \frac{\partial}{\partial x_k}\left[a_{ik}(x)\frac{\partial}{\partial x_i} + a_k(x)\right] + \sum_{k=1}^{m} b_k(x)\frac{\partial}{\partial x_k} + c(x)$$

is a given elliptic operator in Ω, recall that z is the (weak) solution, in the sense of [12], of the problem

$$\text{(II)} \qquad \begin{aligned} Lz &= f(x) \quad \text{in } \Omega, \\ z(x) &= 0 \quad \text{in } \partial\Omega, \end{aligned}$$

if $z \in H_0^1(\Omega)$ and

$$\int_\Omega \left(\sum_{i,k=1}^{m} a_{ik} z_{x_i} h_{x_k} + \sum_{k=1}^{m} a_k z h_{x_k} + \sum_{k=1}^{m} b_k z_{x_k} h + czh - fh \right) dx = 0$$

for every $h \in C_0^1(\Omega)$.

If $L(u)$ is as in (I) (in (II)) for every u in a given class U of functions such that $u(y) \in K$ a.e. in Q (in Ω), where $y \equiv (x,t)$ ($y \equiv x$), with coefficients $a_{ik}(y,u)$, $a_k(y,u)$, $b_k(y,u)$, $c(y,u)$, and if $f(y,u)$ is given in (I) (in (II)) as a right-hand side, then the following uniform parabolic (elliptic) structure hypothesis will be used.

HYPOTHESIS. a_{ik}, a_k, b_k, c, f are Carathéodory functions in $Q \times K$ (in $\Omega \times K$); $a_{ik}(y,\cdot)$, $a_k(y,\cdot)$, $b_k(y,\cdot)$, $c(y,\cdot)$, $f(y,\cdot)$ belong to $C^1(K)$ a.e. in Q (a.e. in Ω) with bounded gradient for almost every $y \in Q$ ($y \in \Omega$) and every $u \in K$ (for every i, k); there exist constants $\alpha > 0$, ω such that $\alpha|\lambda|^2 \leqq \sum_{i,k=1}^m a_{ik}(y,u)\lambda_i\lambda_k \leqq \omega|\lambda|^2$ a.e. in Q (in Ω) for every $u \in K$ and $\lambda \in R^m$; $a_k(u)$, $b_k(u) \in L^{p,q}(Q)$ for every $u \in U$ and k, where $2 < p$, $q \leqq \infty$ and $m/2p + 1/q < 1/2$; $c(u)$, $f(u) \in L^{p',q'}(Q)$ for every $u \in U$, where $1 < q', p' \leqq \infty$ and $m/2p' = 1/q' < 1$ ($b_k(u) \in L^r(\Omega)$, $a_k(u) \in L^r(\Omega)$ (for every k), $c(u) \in L^{r/2}(\Omega)$ for some $r > m$ and for every $u \in U$; $\|a_k(u)\|$, $\|b_k(u)\|$, $\|c(u)\|$

(the norms being computed in the appropriate Lebesgue space) are sufficiently small for every u and k; $f(u) \in L^{\bar r}(\Omega)$, where $\bar r \equiv 2m/(m+2)$ if $m > 2$, $\bar r \equiv 2r/(r+2)$ with $r > 2$ if $m = 2$; $\bar r \equiv 2$ if $m = 1$).

3. Necessary conditions. The following theorem gives a necessary condition for optimal controls of Cauchy–Dirichlet parabolic problems.

THEOREM 1. *Let Ω be connected, K convex in R^s,*

$$U \equiv \{u \in L^\infty(Q) : u(x,t) \in K \text{ a.e. in } Q\},$$

$$L(u)z \equiv z_t - \sum_{i,k=1}^{m} \frac{\partial}{\partial x_k}[a_{ik}(x,t,u(x,t))z_{x_i} + a_k(x,t,u(x,t))z] - \sum_{k=1}^{m} b_k(x,t,u(x,t))z_{x_k} - c(x,t,u(x,t))z.$$

Let $z(u)$ be, for every $u \in U$, the solution, in the sense of [1], *of the problem*

$$L(u)z(u) = f(x,t,u(x,t)) \quad \text{in } Q,$$

$$z(u)(x,0) = z_0(x) \quad \text{in } \Omega, \qquad z(u)(x,t) = 0 \quad \text{in } \partial\Omega \times [0,T].$$

Suppose that u_0 minimizes, in U, $u \to \int_\Omega H(x, z(u)(x,T))\,dx$ and that

(a) *the uniform structure hypothesis holds*;

(b) $z_0 \in L^2(\Omega)$;

(c) *H is a Carathéodory function in $\Omega \times R^1$*;

(d) *there exists a Carathéodory function $\tilde l$ such that $H(x,y_1) \leqq H(x,y_2) + \tilde l(x,y_2)(y_1 - y_2)$ for almost every $x \in \Omega$, every y_1 and y_2, and $|\tilde l(x,y)| \leqq l_0(x) + l_1|y|$ for some $l_0 \in L^2(\Omega)$ and $l_1 \in R^1$*;

(e) *$H(\cdot, \bar y) \in L^1(\Omega)$ for some $\bar y$.*

Then, setting $z^0 \equiv z(u_0)$, $u_0(x,t)$ maximizes in K, a.e. in Q,

$$v \to v \cdot \Big\{ \sum_{i,k=1}^{m} [a_{iku}(x,t,u_0(x,t))z^0_{x_i}(x,t) + a_{ku}(x,t,u_0(x,t))z^0(x,t)]z^*_{x_k}(x,t) - \sum_{k=1}^{m} b_{ku}(x,t,u_0(x,t))z^0_{x_k}(x,t)z^*(x,t) - c_u(x,t,u_0(x,t))z^0(x,t)z^*(x,t) - f_u(x,t,u_0(x,t))z^*(x,t) \Big\},$$

where z^* is the solution, in the sense of [1], *of the problem*

$$L^*(u_0)z^* = 0 \quad \text{in } Q,$$

$$z^*(x,T) = \tilde l(x, z^0(x,T)) \quad \text{in } \Omega, \qquad z^*(x,t) = 0 \quad \text{in } \partial\Omega \times [0,T],$$

$L^*(u_0)$ being the adjoint of $L(u_0)$.

Proof. From [1, Theorem 1, p. 634] it follows that, by (a) and (d), $z(u)$ and z^* are well-defined for every $u \in U$. Moreover, from (d) we see that if $z \in L^2(\Omega)$, then, a.e. in Ω,

$$H(x,\bar y) - \tilde l(x, z(x))(\bar y - z(x)) \leqq H(x, z(x)) \leqq H(x,\bar y) + \tilde l(x,\bar y)(z(x) - \bar y),$$

so $H(\cdot, z(\cdot)) \in L^1(\Omega)$. If everything is well-defined, set

$$(L(u)z, \varphi) \equiv \int_\Omega \left\{ -z\varphi_t + \sum_{i,k=1}^{m} [a_{ik}(u)z_{x_i} + a_k(u)z]\varphi_{x_k} - \sum_{k=1}^{m} b_k(u)z_{x_k}\varphi - c(u)z\varphi \right\} dx\, dt.$$

Note that, for every $u \in U$,

$$L^*(u)z \equiv -z_t - \sum_{i,k=1}^{m} \frac{\partial}{\partial x_i}[a_{ik}(u)z_{x_k} - b_i(u)z] + \sum_{i=1}^{m} a_i(u)z_{x_i} - c(u)z.$$

Set $l^*(x) \equiv \tilde{l}(x, z(u_0)(x, T))$. For every $u \in U$ extend the definition of the coefficients of $L(u)$ in $R^{m+1} - Q$ by setting $L(u) \equiv z_t - \Delta z$ there; moreover, set $f(x, t, u) \equiv 0$ if $(x, t) \notin Q$, for every u, and $l^*(x) = z_0(x) \equiv 0$ if $x \notin \Omega$ (the same notations holding for the original and the extended functions). For every u and $n = 1, 2, \cdots$, let $L^n(u)$ be the operator whose coefficients are the integral averages of the corresponding ones formed with a kernel whose support lies in $\{(x, t) \in R^{m+1} : |x|^2 + t^2 < 1/n^2\}$; let z_0^n, l^n be integral averages of z_0, l^* respectively with a kernel whose support lies in $\{x \in R^m : |x| < 1/n\}$; let $\{\Omega_n\}$ be a sequence of open connected sets exhausting Ω, such that $\bar{\Omega}_n \subset \Omega_{n+1} \subset \Omega$ for every n and $\partial\Omega_n$ is smooth so that any solution of a Cauchy–Dirichlet problem in $Q_n \equiv \Omega_n \times (0, T]$ with $C^\infty(Q_n)$ coefficients and data belongs to $C^1(\bar{Q}_n)$; and finally let d_n be in $C_0^\infty(\Omega_n)$ with $d_n(x) \equiv 1$ if $x \in \Omega_{n-1}$, $0 \leqq d_n(x) \leqq 1$ for every x. (If z_0 has compact support in Ω, define $d_n(x) \equiv 1$ for every x.) Extend (with the same notation) the definition of d_n setting $d_n(x) \equiv 0$ in $\Omega - \Omega_n$. Let $z^n(u)$ be the solution, for every n and u, of

$$L^n(u)z^n(u) = f^n \quad \text{in } Q_n,$$

$$z^n(u)(x, 0) = d_n(x)z_0^n(x) \quad \text{in } \Omega_n, \qquad z^n(u)(x, t) = 0 \quad \text{in } \partial\Omega_n \times [0, T].$$

Set $z^n(u)(x, t) \equiv 0$ if $(x, t) \in Q - \bar{Q}_n$. Let $\varepsilon \in [0, 1]$, $u_0 + v \in U$. Then $u_0 + \varepsilon v \in U$. Set

$$z^\varepsilon \equiv z(u_0 + \varepsilon v),$$

$$a_{ik}^\varepsilon = a_{ik}(u_0 + \varepsilon v), \cdots, f^\varepsilon \equiv f(u_0 + \varepsilon v),$$

$$z^{\varepsilon n} \equiv z^n(u_0 + \varepsilon v),$$

$$L^\varepsilon \equiv L(u_0 + \varepsilon v).$$

Let z^{*n} be the solution, for every n, of

$$L^{*n}(u_0)z^{*n} = 0 \quad \text{in } Q_n,$$

$$z^{*n}(x, T) = d_n(x)l^n(x) \quad \text{in } \Omega_n, \qquad z^{*n}(x, t) = 0 \quad \text{in } \partial\Omega_n \times [0, T].$$

Set $z^{*n}(x, t) \equiv 0$ if $(x, t) \in Q - \bar{Q}_n$. Thus if w_n, w denote, respectively, either $z^n(u)$, $z(u)$, or z^{*n}, z^*, we get (from [1, Theorem 1 and its proof, p. 634]), for every u,

$$(1) \qquad w_n \to w \quad \text{in } L^2(Q), \qquad w_n \rightharpoonup w \quad \text{in } L^2(H_0^1(\Omega)) \text{ and in } L^{2\dot{p}, 2\dot{q}}(Q)$$

for every $\dot{p}$, $\dot{q}$ whose conjugates p, q satisfy $1 < p, q \leqq \infty$, $m/2p + 1/q < 1$.

Since $z^{*n}(x, t) = 0$ if $x \notin \Omega_n$, $z^{*n}(\cdot, t) \in H_0^1(\Omega)$ for every t, and $z^{*n} \in C^1(\bar{Q}_n)$ for every n, we see that the Green's formula (see [1, Formula 2.2, p. 622]) holds true (by density) with z^{*n} as a "test function." Thus, for every n and ε,

$$
\begin{aligned}
(L^0 z^0, z^{*n}) - (L^\varepsilon z^\varepsilon, z^{*n}) &= ((L^0 - L^\varepsilon) z^\varepsilon, z^{*n}) + (L^0(z^0 - z^\varepsilon), z^{*n}) \\
&= \int_Q (f^0 - f^\varepsilon) z^{*n}\, dx\, dt \\
&\quad + \int_\Omega [z^\varepsilon(x, T) - z^0(x, T)]\, d_n(x) l^n(x)\, dx. \qquad (2)
\end{aligned}
$$

Integrating by parts and using Green's formula, for every n, k and ε we see that

$$
\begin{aligned}
(L^0(z^0 - z^\varepsilon), z^{*n}) &= (L^0(z^0 - z^\varepsilon) - L^0(z^{0k} - z^{\varepsilon k}), z^{*n}) + (L^0(z^{0k} - z^{\varepsilon k}), z^{*n}); \\
(L^0(z^{0k} - z^{\varepsilon k}), z^{*n}) &= (z^{0k} - z^{\varepsilon k}, L^{*0} z^{*n}) \\
&\quad - \int_\Omega [z^{0k}(x, T) - z^{\varepsilon k}(x, T)]\, d_n(x) l^n(x)\, dx; \\
(z^{0k} - z^{\varepsilon k}, L^{*0} z^{*n}) &= (z^{0k} - z^{\varepsilon k}, (L^{*0} - L^{*n}) z^{*n}) \\
&\quad + \int_\Omega d_n(x) l^n(x) [z^{0k}(x, T) - z^{\varepsilon k}(x, T)]\, dx.
\end{aligned}
$$

Thus, for every n, k, ε,

$$
(L^0(z^0 - z^\varepsilon), z^{*n}) = (L^0(z^0 - z^{0k} + z^{\varepsilon k} - z^\varepsilon), z^{*n}) + ((L^{*0} - L^{*n}) z^{*n}, z^{0k} - z^{\varepsilon k}). \qquad (3)
$$

Letting $k \to \infty$ in (3) we have, from (1),

$$
(L^0(z^0 - z^\varepsilon), z^{*n}) = ((L^{*0} - L^{*n}) z^{*n}, z^0 - z^\varepsilon) \quad \text{for every } n \text{ and } \varepsilon;
$$

and letting $n \to \infty$ in the last formula, remembering (i), (ii), (d), observing that $d_n(x) \to 1$ for every $x \in \Omega$ and that u_0 minimizes, we get, for every ε,

$$
((L^0 - L^\varepsilon) z^\varepsilon, z^*) \geqq \int_Q (f^0 - f^\varepsilon) z^*\, dx\, dt;
$$

that is,

$$
\begin{aligned}
\frac{1}{\varepsilon} \int_Q \Big\{ &\sum_{i,k=1}^m [(a_{ik}^\varepsilon - a_{ik}^0) z_{x_i}^\varepsilon + (a_k^\varepsilon - a_k^0) z^\varepsilon] z_{x_k}^* - \sum_{k=1}^m (b_k^\varepsilon - b_k^0) z_{x_k}^\varepsilon z^* \\
&- (c^\varepsilon - c^0) z^* z^\varepsilon - (f^\varepsilon - f^0) z^* \Big\}\, dx\, dt \leqq 0 \quad \text{for every } \varepsilon \in (0, 1]. \qquad (4)
\end{aligned}
$$

Since (see [1]) $z_x^\varepsilon \rightharpoonup z_x^0$ in $L^2(Q)$ and, by (a), the incremental ratios of the coefficients in (4) converge uniformly as $\varepsilon \to 0+$, we have

$$
\begin{aligned}
\int_Q \Big\{ &\sum_{i,k=1}^m [a_{iku}(u_0) z_{x_i}^0 + a_{ku}(u_0) z^0] z_{x_k}^* - \sum_{k=1}^m b_{ku}(u_0) z_{x_k}^0 z^* \\
&- c_u(u_0) z^0 z^* - f_u(u_0) z^* \Big\} \cdot (v - u_0)\, dx\, dt \leqq 0 \quad \text{for every } v \in U. \qquad (5)
\end{aligned}
$$

(Equation (5) is a necessary condition in integral form for u_0, which maximizes,

in U, the linear functional

$$u \to \int_Q \left\{ \sum_{i,k=1}^{m} [a_{iku}(u_0) z^0_{x_i} z^*_{x_k} + \cdots - f_u(u_0) z^*] \right\} \cdot u \, dx \, dt.)$$

Given $v \in K$, $\bar{y} \equiv (\bar{x}, \bar{t}) \in Q$, let n be a positive integer such that

$$S_n \equiv \{y \in Q : |y - \bar{y}| \leqq 1/n\} \subset Q.$$

Set

$$w_n(x,t) \equiv \begin{cases} v & \text{if } (x,t) \in S_n, \\ u_0(x,t) & \text{if } (x,t) \in Q - S_n. \end{cases}$$

Then $w_n \in U$ for every n; therefore, from (5),

$$\int_{S_n} \left\{ \sum_{i,k=1}^{m} [a_{iku}(u_0) z^0_{xi} + a_{ku}(u_0) z^0] z^*_{x_k} - \sum_{k=1}^{m} b_{ku}(u_0) z^0_{x_k} z^* \right. \tag{6}$$
$$\left. - c_u(u_0) z^0 z^* - f_u(u_0) z^* \right\} \cdot (v - u_0) \, dx \, dt \leqq 0 \quad \text{for every } n.$$

Dividing (6) by meas. S_n as $n \to \infty$ we have (K being separable)

$$\left\{ \sum_{i,k=1}^{m} [a_{iku}(x,t,u_0(x,t)) z^0_{x_i}(x,t) + a_{ku}(x,t,u_0(x,t)) z^0(x,t)] z^*_{x_k}(x,t) \right.$$
$$- \sum_{k=1}^{m} b_{ku}(x,t,u_0(x,t)) z^0_{x_k}(x,t) z^*(x,t) - c_u(x,t,u_0(x,t)) z^0(x,t) z^*(x,t)$$
$$\left. - f_u(x,t,u_0(x,t)) z^*(x,t) \right\} \cdot (v - u_0(x,t)) \leqq 0$$

a.e. in Q and for every $v \in K$. This completes the proof.

Remark 1. An existence theorem concerning the above control problems is proved in [13, Theorem 3] assuming the structure hypotheses, and (more restrictively than here) that $U \equiv \{u \in L^\infty(\Omega) : u(x) \in K \text{ q.o. in } \Omega\}$, that (a_{ik}) is symmetric and independent on t, that K is compact, that the range of $(a_{11}(x, \cdot), \cdots, a_{mm}(x, \cdot))$ is, a.e. in Ω, maximal (if $m = 1$, this maximality hypothesis can be replaced by convexity of this range), that $a_k = 0 = b_k = c$ for every k, and that f is independent on the control. In such a case it is easy to show that, if the coefficients are differentiable with respect to u as assumed in Theorem 1, the necessary condition for u_0 is the following:

$$\sum_{i,k=1}^{m} a_{iku}(x, u_0(x)) \int_0^T z^0_{x_i}(x,t) z^*_{x_k}(x,t) \, dt \cdot (v - u_0(x)) \leqq 0$$

a.e. in Ω and for every $v \in K$.

If the coefficients a_{ik}, a_k are smooth and independent on the control, and if $c = 0$, from [6, Theorem 2, p. 202] it follows that u_0 maximizes (a.e. in Q) in K:

$$v \to \left[\sum_{k=1}^{m} b_k(x,t,v) z^0_{x_k}(x,t) + f(x,t,v) \right] z^*(x,t)$$

(without differentiability hypotheses of coefficients with respect to u). Other existence theorems for Cauchy–Dirichlet problems are proved in [8].

The following theorem gives a necessary condition (analogous to Theorem 1) for optimal controls of Dirichlet problems.

THEOREM 2. *Let K be a convex set in R^s,*

$$U \equiv \{u \in L^\infty(\Omega) : u(x) \in K \text{ a.e. in } \Omega\},$$

$$L(u)z \equiv -\sum_{i,k=1}^{m} \frac{\partial}{\partial x_k}[a_{ik}(x, u(x))z_{x_i} + a_k(x, u(x))z] + \sum_{k=1}^{m} b_k(x, u(x))z_{x_k} + c(x, u(x))z.$$

Let $z(u)$ be the solution, for every $u \in U$, in the sense of [12], *of the problem*

$$L(u)z(u) = f(x, u(x)) \quad \text{in } \Omega,$$

$$z(u)(x) = 0 \quad \text{in } \partial\Omega.$$

Suppose that u_0 minimizes, in U, $u \to \int_\Omega H(x, z(u)(x))\,dx$, and that

(a) *the uniform structure hypothesis holds;*

(b) *H is a Carathéodory function in $\Omega \times R^1$, and there exists $\tilde{l}$ of Carathéodory in $\Omega \times R^1$ such that*

$$H(x, y_1) \leqq H(x, y_2) + \tilde{l}(x, y_2)(y_1 - y_2)$$

a.e. in Ω and for every y_1, y_2, where $|\tilde{l}(x, y)| \leqq \tilde{l}_0(x) + l_1|y|^{2/\bar{r}}$ for some $\tilde{l}_0 \in L^{\bar{r}}(\Omega)$ and $l_1 \in R^1$; $H(\cdot, \bar{y}) \in L^1(\Omega)$ for some $\bar{y}$.

Then $u_0(x)$ maximizes, a.e. in Ω,

$$v \to v \cdot \left\{ \sum_{i,k=1}^{m} [a_{iku}(x, u_0(x))z^0_{x_i}(x) + a_{ku}(x, u_0(x))z^0(x)]z^*_{x_k}(x) + \sum_{k=1}^{m} b_{ku}(x, u_0(x))z^0_{x_k}(x)z^*(x) + c_u(x, u_0(x))z^0(x)z^*(x) - f_u(x, u_0(x))z^*(x) \right\},$$

with z^* being the solution, in the sense of [12], *of the problem*

$$L^*(u_0)z^* = \tilde{l}(x, z(u_0)(x)) \quad \text{in } \Omega,$$

$$z^*(x) = 0 \quad \text{in } \partial\Omega,$$

where $L^*(u_0)$ denotes the adjoint of $L(u_0)$.

Proof. From (a) and (b), $z(u)$ and z^* are well-defined for every $u \in U$ (see [12]). Suppose $\varepsilon \in (0, 1]$, $u + v \in U$. From the definition of z^*, we have, for every $\varphi \in H_0^1(\Omega)$,

$$\int_\Omega \left\{ \sum_{i,k=1}^{m} [a_{ik}(u_0)z^*_{x_k} + b_i(u_0)z^*]\varphi_{x_i} + \sum_{k=1}^{m} a_k(u_0)z^*_{x_k}\varphi + c(u_0)z^*\varphi \right\} dx = \int_\Omega l_0\varphi\,dx,$$

where $l_0(x) \equiv \tilde{l}(x, z(u_0)(x))$. Thus, using notation similar to that in the proof of Theorem 1, and remembering (b), we get, for every ε,

$$\int_\Omega \Big\{ \sum_{i,k=1}^{m} [a_{ik}(u_0)z^*_{x_k} + b_i(u_0)z^*](z^\varepsilon_{x_i} - z^0_{x_i}) + \sum_{k=1}^{m} a_k(u_0)z^*_{x_k}(z^\varepsilon - z^0) + c(u_0)z^*(z^\varepsilon - z^0)\Big\}\, dx \tag{7}$$

$$= \int_\Omega l_0(z^\varepsilon - z^0)\, dx$$

$$= (z^*, L^0(z^\varepsilon - z^0)) = ((L^0 - L^\varepsilon)z^\varepsilon, z^*) + \int_\Omega (f^\varepsilon - f^0)z^*\, dx.$$

From (7), since (b) holds, we get, for every ε,

$$\frac{1}{\varepsilon}\int_\Omega \Big\{ \sum_{i,k=1}^{m} [(a^\varepsilon_{ik} - a^0_{ik})z^\varepsilon_{x_i} + (a^\varepsilon_k - a^0_k)z^\varepsilon]z^*_{x_k} + \sum_{k=1}^{m} (b^\varepsilon_k - b^0_k)z^\varepsilon_{x_k}z^* + (c^\varepsilon - c^0)z^\varepsilon z^* - (f^\varepsilon - f^0)z^*\Big\}\, dx \leqq 0. \tag{8}$$

As in the proof of Theorem 1 the conclusion follows from (8).

Remark 2. Other existence and uniqueness theorems for $z(u)$ and z^* are proved in [9]. An existence theorem for the above control problem is proved in [13, Theorem 4], assuming the structure hypotheses and that $a_{ik} = a_{ki}$, $b_i = a_i = c = 0$ for every i, f is independent on u, and that, a.e. in Ω, the range of $(a_{11}(x, \cdot), \cdots, a_{mm}(x, \cdot))$ is maximal.

4. Uniqueness and "bang-bang" theorems. The following lemma is a form of the maximum principle for solutions of parabolic problems.

LEMMA. *Let Ω be connected, L a parabolic operator, $f \in L^{p,q}(Q)$ and $u_0 \in L^2(\Omega)$ such that the hypotheses of* [1, *Theorem* 1, *p.* 634] *are satisfied. If u is the solution, in the sense of* [1], *of*

$$Lu = f \quad \text{in } Q,$$

$$u = u_0 \quad \text{in } \Omega, \qquad u = 0 \quad \text{in } \partial\Omega \times [0, T],$$

then $u(x, t) > 0$ in Q if $f \geqq 0$, $u_0 \geqq 0$, $u_0 \neq 0$.

Proof. If u_1 solves $Lu_1 = 0$ in Q, $u_1 = u_0$ in Ω, $u_1 = 0$ in $\partial\Omega \times [0, T]$, and u_2 solves $Lu_2 = f$ in Q, $u_2 = 0$ in Ω, $u_2 = 0$ in $\partial\Omega \times [0, T]$, we see that $u = u_1 + u_2$, $u_i \geqq 0$ in Q $(i = 1, 2)$ (see [1, Theorem 1, p. 634 and Theorem 9, p. 671]). If $u_1(x, t) = 0$ in Q, from

$$\lim_{t \to 0+} \int_\Omega u_1(x, t)\psi(x)\, dx = \int_\Omega u_0(x)\psi(x)\, dx \quad \text{for every } \psi \in C_0^1(\Omega), \tag{9}$$

we deduce $u_0 = 0$; so there exists $(x_0, t_0) \in Q$ such that $u_1(x_0, t_0) > 0$. From Harnack's inequality [1, Theorem H, p. 618] we see (Ω being open connected) that $u_1(x, t) > 0$ if $x \in \Omega$ and $t_0 < t \leqq T$. If there exist $\bar{s} \in (0, t_0]$, $\bar{y} \in \Omega$ such that $u_1(\bar{y}, \bar{s}) = 0$, again from Harnack's inequality we get $u_1(y, s) = 0$ in $\Omega \times (0, \bar{s})$; therefore, from (9), $u_0 = 0$, which is absurd; so $u_1(x, t) > 0$ in Q.

From Theorems 1 and 2 follow uniqueness criteria of optimal controls.

COROLLARY 1. *Suppose that (using notations of Theorem* 1 *and its proof)*:

(a) *the hypotheses of Theorem* 1 *hold true*;

(b) $a_{ik}(x,t,\cdot)$, $a_k(x,t,\cdot)$, $b_k(x,t,\cdot)$ *are linear a.e. in* Q *and for every* i,k;

(c) *a.e. in* Q *one of the following four hypotheses holds*:

(i) $l^* \geqq 0$ *and* $z_0 \geqq 0$ $(\leqq 0)$, $l^* \neq 0$, $z_0 \neq 0$; $f \geqq 0$ $(\leqq 0)$; $c(x,t,\cdot)$ *convex (concave),* $f(x,t,\cdot)$ *convex, one at least strictly*;

(ii) $l^* \leqq 0$ *and* $z_0 \geqq 0$ $(\leqq 0)$, $l^* \neq 0$, $z_0 \neq 0$; $f \geqq 0$ $(\leqq 0)$; $c(x,t,\cdot)$ *concave convex),* $f(x,t,\cdot)$ *concave, one at least strictly.*

Then there exists at most one optimal control for the parabolic problem of Theorem 1.

Proof. Let u_0 be an optimal control, z^0 the corresponding solution, and set

$$B(x,t,u) \equiv \sum_{i,k=1}^{m} [a_{ik}(x,t,u)z^0_{x_k}(x,t) + a_k(x,t,u)z^0(x,t)]z^*_{x_k}(x,t)$$
$$- \sum_{k=1}^{m} b_k(x,t,u)z^0_{x_k}(x,t)z^*(x,t) - c(x,t,u)z^0(x,t)z^*(x,t)$$
$$- f(x,t,u)z^*(x,t).$$

From (b) and (c), $B(x,t,\cdot)$ is strictly concave a.e. in Q: for if, for example, $l^* \geqq 0$, $z_0 \geqq 0$, as in (i), then $z^0 > 0$ in Q and $z^* > 0$ in Q, because of the lemma.

u_0 is an optimal control; therefore, from Theorem 1, $u_0(x,t)$ is, a.e. in Q, the unique maximum, in K, of $v \to B(x,t,v)$.

In a completely analogous way, from the maximum principle for elliptic equations (see [4], [5]), we have the following corollary.

COROLLARY 2. *Suppose that (using notations of Theorem* 2 *and its proof)*:

(a) *the hypotheses of Theorem* 2 *hold, and* Ω *is connected*;

(b) $a_{ik}(x,\cdot)$, $a_k(x,\cdot)$, $b_k(x,\cdot)$ *are linear a.e. in* Ω *and for every* i,k;

(c) *for every* $u \in U$, $c(u) \geqq \sum_{k=1}^{m} [a_k(u)]_{x_k}$, $c(u) \geqq \sum_{k=1}^{m} [b_k(u)]_{x_k}$ *(in the sense of distributions)*;

(d) *a.e. in* Ω *one of the following four hypotheses holds*:

(i) $l_0 \geqq 0, f \geqq 0$ $(\leqq 0)$, $l_0 \neq 0, f \neq 0$; $c(x,t,\cdot)$ *convex (concave),* $f(x,t,\cdot)$ *convex, one at least strictly, a.e. in* Q;

(ii) $l_0 \leqq 0, f \geqq 0$ $(\leqq 0)$, $l_0 \neq 0, f \neq 0$; $c(x,t,\cdot)$ *concave (convex),* $f(x,t,\cdot)$ *concave, one at least strictly, a.e. in* Q.

Then there exists at most one optimal control for the elliptic control problem of Theorem 2.

If the second and first order coefficients do not depend on controls, every optimal control is "bang-bang" (meaning that its range is contained in ∂K) in the hypotheses of the next corollary.

COROLLARY 3. *Suppose that (using notations of Theorem* 1 *and its proof)*:

(a) *the hypotheses of Theorem* 1 *hold*;

(b) a_{ik}, a_k, b_k *do not depend on* u *for every* i,k;

(c) $l^*(x) \geqq 0$, *or* $l^*(x) \leqq 0$ *a.e. in* Ω, $l^* \neq 0$;

(d) $z_0(x) \geqq 0$ *a.e. in* $\Omega, f(x,t,u) \geqq 0, f_u(x,t,u) \neq 0, f_{u_j}(x,t,u) \geqq 0, c_{u_j}(x,t,u) \geqq 0$, *for every* $j = 1, \cdots, s$ *a.e. in* Q *and for every* $u \in K$; *or*

(d') $|c_u(x,t,u)y + f_u(x,t,u)| > 0$ *a.e. in* Q *for every* $u \in U$ *and for every* $y \in z^0(Q)$.

Then if u_0 *is an optimal control for the parabolic problem, a.e. in* Q, *we have* $u_0(x,t) \in \partial K$.

Proof. Let u_0 be an optimal control with corresponding solution z^0, and set

$$B(x,t) \equiv c_u(x,t,u_0(x,t))z^0(x,t) + f_u(x,t,u_0(x,t)).$$

From the maximum principle for parabolic equations (see Lemma), (c) implies that $z^*(x,t) \neq 0$ in Q. Therefore, from (b) and Theorem 1, $u_0(x,t)$ is an absolute extremum, in K, of $v \to B(x,t)\cdot v$ a.e. in Q. From (d), $z^0(Q) \subset [0,+\infty)$ (see [1]), and, for every $u \in K$, $y \to |c_u(x,t,u)y + f_u(x,t,u)|^2$ is (a.e. in Q) nondecreasing in $[0,+\infty)$. Therefore, again from (c), (d') follows from (d). Suppose, for example, that $u_0(x,t)$ minimizes in K, $v \to B(x,t)\cdot v$ (a.e. in Q) (the proof is entirely analogous if maximizing). Set $E_0 \equiv \{(x,t) \in Q : u_0(x,t) \in K^\circ\}$. If meas. $E_0 > 0$ and we set $d(x,t) \equiv d(u_0(x,t), \partial K)$, in E_0 we get $u_0(x,t) + w\, d(x,t) \in K$ for every $w \in R^s$ such that $|w| \leqq 1$, so $B(x,t)\cdot w\, d(x,t) \geqq 0$ a.e. in E_0 and for the same w's, which is absurd. Therefore meas. $E_0 = 0$.

Entirely analogous is the proof of the following corollary.

COROLLARY 4. *Suppose that (using notations of Theorem 2 and its proof):*

(a) *the hypotheses of Theorem 2 hold;*

(b) a_{ik}, a_k, b_k *are independent on* u *for every* i,k;

(c) *for every* $u \in U$, $c(u) \geqq \sum_{k=1}^m b_{kx_k}$ *(in the distributional sense);* $l_0(x) \geqq 0$, *or* $l_0(x) \leqq 0$, *a.e. in* Ω; $l_0 \neq 0$;

(d) $f(x,u) \geqq 0$, $f_u(x,u) \neq 0$. $f_{u_j}(x,u) \geqq 0$, $c_{u_j}(x,u) \geqq 0$ *for every* $j = 1, \cdots, s$, *a.e. in* Ω *for every* $u \in K$; $c(u) \geqq \sum_{k=1}^m a_{kx_k}$ *(in the sense of distributions) for every* $u \in K$;

or

(d') *a.e. in* Ω, $|c_u(x,u)y + f_u(x,u)| > 0$ *for every* $u \in K$ *and for every* $y \in z^0(\Omega)$.

Then, if u_0 *is an optimal control for the elliptic problem,* $u_0(x) \in \partial K$ *a.e. in* Ω.

Remark 3. If f only depends on u, bang-bang theorems are known: see [7]. Clearly different hypotheses in Corollaries 3 and 4 could be chosen to give the same conclusions.

5. Stability theorems. From the physical meaning of the present control problems the question arises in a natural way (see [2]) as to whether these problems are "well-posed" in a variational sense. About the parabolic problem, the following theorem shows that if the hypotheses of Theorem 3 in [13] hold, so obtaining existence of optimal controls of the approximating and the original problem, then the optimal states of the approximating problems converge to an optimal state of the original problem, and the minima of the approximate problems converge to the minimum of the original one, if the functional is continuous.

THEOREM 3. *Let* Ω *be connected,* K *a nonvoid compact set in* R^s, $U \equiv \{u \in L^\infty(\Omega) : u(x) \in K$ *a.e. in* $\Omega\}$, $0 < t_0$, $\bar{t} \leqq T$, $x_0 \in \Omega$. *For every* $u \in U$ *and* $n = 0, 1, 2, \cdots$, *let* $z_n(u)$ *be the solution, in the sense of* [1], *of*

$$z_{nt}(u) - \sum_{i,k=1}^m \frac{\partial}{\partial x_i}[a_{nik}(x, u(x))z_{nx_k}(u)] = 0 \quad \text{in } Q;$$

$$z_n(u)(x,0) = z^0(x) \quad \text{in } \Omega, \qquad z_n(u)(x,t) = 0 \quad \text{in } \partial\Omega \times [0,T].$$

If u_n minimizes, in U, for $n = 1, 2, \cdots$,

$$u \to G_n(u) \equiv F\left(z_n(u)(x_0, t_0), \int_\Omega h(x, z_n(u)(x, \bar{t}))\, dx\right),$$

then there exists u_0 which minimizes G_0 in U, such that (for some subsequence) $z_n(u_n)(x, t) \to z_0(u_0)(x, t)$ uniformly on compact sets in Q, $z_n(u_n) \rightharpoonup z_0(u_0)$ in $L^2(H_0^1(\Omega))$, $\min G_n(U) \to \min G_0(U)$, if

(a) *$a_{nik} = a_{nki}$ are Carathéodory functions in $\Omega \times K$, for every $i, k = 1, \cdots, m$, and for every $n = 0, 1, 2, \cdots$; there exist constants $\alpha > 0$, ω such that, for every $\lambda \in R^m$, $u \in K$, $n = 0, 1, 2, \cdots$, a.e. in Ω, we have*

$$\alpha|\lambda|^2 \leqq \sum_{i,k=1}^{m} a_{nik}(x, u)\lambda_i\lambda_k \leqq \omega|\lambda|^2;$$

setting $I_{kk} \equiv [\alpha, \omega]$,

$I_{ik} \equiv [(\alpha - \omega)/2, (\omega - \alpha)/2]$ if $i \neq k$, $P \equiv I_{11} \times \cdots \times I_{mm}$, we have

$$\{(a_{n\,1\,1}(x, u), \cdots, a_{n\,m\,m}(x, u)) : u \in K\} = P$$

a.e. in Ω and for $n = 0, 1, 2, \cdots$; $z^0 \in H_0^1(\Omega)$;

(b) *$a_{n\,i\,k}(x, u) \to a_{0\,i\,k}(x, u)$ for every i, k, $u \in K$ and a.e. in Ω;*

(c) *h is a Carathéodory function in $\Omega \times R^1$; $|h(x, v)| \leqq H(x)$ for every v, a.e. in Ω, for some $H \in L^1(\Omega)$; $F : R^2 \to R^1$ is continuous.*

Proof. From [13, Theorem 3] follows, by (a), (b), (c), the existence of u_n, $n = 0, 1, 2, \cdots$. Thus $G_n(u_n) \leqq G_n(u)$ for every n and u. From (a) and (b), for some subsequence $z_n(u)(x, t) \to z_0(u)(x, t)$ uniformly on compact sets of Q, for every $u \in U$ (see [1]). Moreover (see [13]) there exists $\bar{z}$ such that (for some subsequence) $z_n(u_n)(x, t) \to \bar{z}(x, t)$ uniformly on compact sets of Q, $z_n(u_n) \rightharpoonup \bar{z}$ in $L^2(H_0^1(\Omega))$; but from results of [10], [11] and from [3, Corollary 5.1′], there exists $u_0 \in U$ such that

$$\bar{z}_t - \sum_{i,k=1}^{m} \frac{\partial}{\partial x_i}(a_{0ik}(x, u_0(x))\bar{z}_{x_k}) = 0 \quad \text{in } Q,$$

$$\bar{z}(x, 0) = z^0(x) \quad \text{in } \Omega, \qquad \bar{z}(x, t) = 0 \quad \text{in } \partial\Omega \times [0, T].$$

Therefore $\bar{z} = z_0(u_0)$, and, from (c), $G_n(u_n) \to G_0(u_0)$ (see proof of Theorem 3 of [13]); hence $G_0(u_0) \leqq G_0(u)$ for every $u \in U$, since, from (c) again, $G_n(u) \to G_0(u)$ for every $u \in U$.

A "stability" theorem without the maximality hypothesis of $\{(a_{n\,1\,1}(x, u), \cdots, a_{n\,m\,m}(x, u) : u \in K\}$ when $m \equiv 1$ now follows.

THEOREM 4. *Let K be a nonvoid compact set in R^s, $U \equiv \{u \in L^\infty(\Omega) : u(x) \in K$ a.e. in $\Omega\}$, $0 < t_0, \bar{t} \leqq T$, $x_0 \in \Omega$. Let $z_n(u)$ be, for every $u \in U$ and $n = 0, 1, 2, \cdots$, the solution (in the sense of [1]) of*

$$z_{nt}(u) - \frac{\partial}{\partial x}[a_n(x, u(x))z_{nx}(u)] = 0 \quad \text{in } Q;$$

$$z_n(u)(x, 0) = z^0(x) \quad \text{in } \Omega, \qquad z_n(u)(x, t) = 0 \quad \text{in } \partial\Omega \times [0, T].$$

If u_n minimizes, in U, for $n = 1, 2, \cdots$,

$$u \to G_n(u) \equiv F\left(z_n(u)(x_0, t_0), \int_\Omega h(x, z_n(u)(x, \bar{t}))\, dx\right),$$

then there exists u_0 *which minimizes* G_0 *in* U, *such that, for some subsequence,* $z_n(u_n)(x,t) \to z_0(u_0)(x,t)$ *uniformly on compact sets of* Q, $z_n(u_n) \rightharpoonup z_0(u_0)$ *in* $L^2(H_0^1(\Omega))$, $\min G_n(U) \to \min G_0(U)$, *if*

(a) a_n, $n = 0, 1, 2, \cdots$, *is a Carathéodory function in* $\Omega \times K$; $z^0 \in H_0^1(\Omega)$;

(b) *there exist constants* $\alpha > 0, \omega$, *such that* $a_n(x,K) \subset [\alpha, \omega]$ *for every* $n = 0, 1, 2, \cdots$, *a.e. in* Ω; $a_n(x,K) \equiv [p_n(x), q_n(x)]$, $n = 0, 1, 2, \cdots$, *and* $1/p_n \rightharpoonup 1/p_0$, $1/q_n \rightharpoonup 1/q_0$ *both in* $L^\infty(\Omega)$;

(c) *hypothesis* (c) *of Theorem* 3 *holds.*

Proof. From (a), (b), (c) we deduce (see [13, Theorem 5]) the existence of u_n, $n = 0, 1, 2, \cdots$. Let g_n belong to $L^\infty(\Omega)$ such that, for every $n = 1, 2, \cdots$, $p_n(x) \leqq g_n(x) \leqq q_n(x)$ a.e. in Ω. If $1/g_n \rightharpoonup 1/g_0$ in $L^\infty(\Omega)$, we see that, a.e. in Ω, $p_0(x) \leqq g_0(x) \leqq q_0(x)$, by (b). Therefore, from [10, Proposition 6, p. 596] there exists, for every $u \in U$, a subsequence such that $z_n(u)(x,t) \to z_0(u)(x,t)$ uniformly on compact sets of Q. From [3, Corollary 5.1'] there exists $u_0 \in U$ such that (for some subsequence) $z_n(u_n)(x,t) \to z_0(u_0)(x,t)$ uniformly on compact sets of Q, and $z_n(u_n) \rightharpoonup z_0(u_0)$ in $L^2(H_0^1(\Omega))$. Therefore $G_n(u_n) \to G_0(u_0)$, and so $G_0(u_0) \leqq G_0(u)$ for every $u \in U$.

Remark 4. In the elliptic case we have two stability theorems completely analogous to Theorems 3 and 4 (whose statements are omitted) for functionals of the form (see [13]) $u \to F(z(u)(x_0), \int_\Omega h(x, z(u)(x))\,dx)$.

From Theorems 3 and 4 we see that the corresponding control problems are "well-posed" in a weak sense, because, generally speaking, it is false that $u_n \rightharpoonup u_0$. A class of parabolic control problems is shown in the next theorem (we omit the completely analogous statement of the elliptic case) which are "well-posed" in a strong sense, because one obtains the convergence of the optimal states and of the minima, and moreover the strong convergence of optimal controls of approximate problems to an optimal control of the original problems.

THEOREM 5. *Let* Ω *be connected,* $K \subset R^s$, $U \equiv \{u \in L^\infty(Q) : u(x,t) \in K \text{ a.e. in } Q\}$. *Let* $z_n(u)$ *be the solution, in the sense of* [1], *for every* $n = 0, 1, 2, \cdots$, *and* $u \in U$, *of*

$$z_{nt}(u) - \sum_{i,k=1}^{m} \frac{\partial}{\partial x_k}[a_{nik} z_n(u)_{x_k} + a_{nk} z_n(u)] - \sum_{k=1}^{m} b_{nk} z_{nx_k}(u) - c_n(u) z_n(u) = f_n(u) \quad \text{in } Q,$$

$$z_n(u)(x,0) = z^0(x) \quad \text{in } \Omega, \qquad z_n(u)(x,t) = 0 \quad \text{in } \partial\Omega \times [0,T].$$

If u_n *is a minimizing control, in* U, *for* $u \to G_n(u) = \int_\Omega H_n(x, z_n(u)(x,T))\,dx$, *then there exists* u_0 *which minimizes* G_0 *in* U, *such that (for some subsequence)* $z_n(u_n)(x,t) \to z_0(u_0)(x,t)$ *uniformly on compact sets of* Q, $z_n(u_n) \rightharpoonup z_0(u_0)$ *in* $L^2(H_0^1(\Omega))$, $\min G_n(U) \to \min G_0(U)$, $u_n \to u_0$ *in every* $L^p(Q)$, $1 \leqq p < \infty$, *if*

(a) *for every* $n = 0, 1, 2, \cdots$, *the hypotheses of Theorem* 1 *hold for the corresponding problem, where* α, ω, p, q *are independent on* n; $|c_n(x,t,u)| \leqq \bar{c}_n(x,t), |f_n(x,t,u)| \leqq \bar{f}_n(x,t)$ *a.e. in* Q, *for every* $u \in K$, *for every* $n = 0, 1, 2, \cdots$, *and for some* $\bar{c}_n, \bar{f}_n \in L^{p,q}(Q)$;

(b) $a_{nik} \to a_{0ik}$ *in* $L^1(Q)$, $a_{nk} \rightharpoonup a_{0k}$ *in* $L^{p,q}(Q)$, $b_{nk} \to b_{0k}$ *in* $L^{p,q}(Q)$ *for every* i, k; $c_n(\cdot, u) \to c_0(\cdot, u)$ *in* $L^2(Q)$, $f_n(\cdot, u) \to f_0(\cdot, u)$ *in* $L^1(Q)$, *both uniformly with respect to* $u \in K$; $H_n(\cdot, y) \to H_0(\cdot, y)$ *in* $L^1(Q)$ *uniformly in* y; $|H_0(x,y)| \leqq h(x)$ *a.e. in* Ω *and for every* y, *with some* $h \in L^1(\Omega)$;

(c) *for every* $n = 0, 1, 2, \cdots$, *the hypotheses of Corollary* 3 *hold for the corresponding problem*;

(d) $\{(c_n(x,t,u), f_n(x,t,u)) : u \in K\}$ *is a convex set a.e. in* Q *and for every* $n \geqq 1$;

(e) $c_0(x,t,\cdot), f_0(x,t,\cdot)$ *are linear*;

(f) $K \equiv \{v \in R^s : |v| \leqq N\}$ *for some* $N > 0$.

Proof. Given $n \geqq 1$, let $\{\bar{u}_k\}$ be a minimizing sequence, in U, for G_n. From (a) there exists a subsequence such that $(c_n(\bar{u}_k), f_n(\bar{u}_k)) \rightharpoonup (\bar{c}_n, \bar{f}_n)$ in $L^{p,q}(Q)$. From (d) and (f) and from [3, Corollary 5.1'] there exists $u_n \in U$ such that $(\bar{c}_n, \bar{f}_n) = (c_n(u_n), f_n(u_n))$, and (for some subsequence) $z_n(\bar{u}_k)(x,t) \to z_n(u_n)(x,t)$ uniformly on compact sets of Q, and $z_n(\bar{u}_k) \to z_n(u_n)$ in $L^2(H_0^1(\Omega))$. The existence of the minimum of $G_n(U)$ is therefore proved if $n \geqq 1$. For every $u \in U$ there exists, by (b), a subsequence such that $z_n(u)(x,t) \to z_0(u)(x,t)$ uniformly on compact sets of Q; therefore (from (b)) $G_n(u) \to G_0(u)$ for every $u \in U$. From (a) and (b) there exists $\bar{z}$ such that (see [1]) $z_n(u_n)(x,t) \to \bar{z}(x,t)$ uniformly on compact sets of Q, and $z_n(u_n) \rightharpoonup \bar{z}$ in $L^2(H_0^1(\Omega))$ (for some subsequence). By (f), if $1 < p < \infty$, there exists $u_0 \in U$ such that (for some subsequence) $u_n \rightharpoonup u_0$ in $L^p(Q)$. If $\varphi \in C_0^1(Q^\circ)$,

$$\int_Q [c_n(u_n)z_n(u_n) - c_0(u_0)\bar{z}]\varphi \, dx \, dt = \int_Q [c_n(u_n) - c_0(u_n)]z_n(u_n)\varphi \, dx \, dt$$

$$+ \int_Q c_0(u_n)[z_n(u_n) - \bar{z}]\varphi \, dx \, dt$$

$$+ \int_Q [c_0(u_n) - c_0(u_0)]\bar{z}\varphi \, dx \, dt \to 0,$$

the first term on the right-hand side from (b) (because $\sup_n \|z_n(u_n)\|_{L^2(Q)} < +\infty$: see [1]), the third one from (e). Similarly (for the same φ's) $\int_Q f_n(u_n)\varphi \, dx \, dt \to \int_Q f_0(u_0)\varphi \, dx \, dt$. Thus $\bar{z} = z_0(u_0)$; therefore u_0 minimizes, in U, G_0 (by (b)). From Corollary 3, $u_n(x,t) \in \partial K$ a.e. in Ω and for every $n = 0, 1\,2, \cdots$. Since $L^p(Q)$, $1 < p < \infty$, is uniformly convex, from (f) we see that $u_n \to u_0$ in every such $L^p(Q)$.

Remark 5. In the above proof, one obtains $u_n \rightharpoonup u_0$ in $L^p(Q)$, $1 \leqq p < \infty$, without the hypothesis (f) about K, thus showing another type of "stability" of these problems.

The next theorem shows that limits of nearly optimal states of approximate problems are optimal states of the original problem, under substantially more general hypotheses than in Theorem 3 (actually existence of optimal states and maximality hypotheses as (a) of Theorem 3 are not required for the approximate problems). Obviously a completely analogous theorem holds in the elliptic control problem.

THEOREM 6. *Let* Ω, T, Q, K, U, t_0, $\tilde{t}$, x_0, a_{nik}, z^0, $z_n(u)$, G_n, F, P *be defined as in Theorem* 3. *Suppose that* $\varepsilon_n > 0$ *for every* $n = 1, 2, \cdots$, *and* $\varepsilon_n \to 0$. *If* $u_n \in U$ *such that* $G_n(u_n) < \varepsilon_n + \inf G_n(U_n)$, $n = 1, 2, \cdots$, *then there exists* $u_0 \in U$ *minimizing* G_0 *in* U, *such that* $z_n(u_n)(x,t) \to z_0(u_0)(x,t)$ *uniformly on compact sets in* Q, $z_n(u_n) \rightharpoonup z_0(u_0)$ *in* $L^2(H_0^1(\Omega))$, $\inf G_n(U) \to \min G_0(U)$, *if*

(a) $a_{nik} = a_{nki}$ *are Carathéodory functions on* $\Omega \times K$ *for every* i, k *and* $n = 0, 1, 2, \cdots$; *there exist constants* $\alpha > 0$, ω *such that, for every* $\lambda \in R^m$,

$u \in K$, $n = 0, 1, 2, \cdots$, *a.e. in* Ω *we have* $\alpha|\lambda|^2 \leqq \sum_{i,k=1}^{m} a_{nik}(x, u)\lambda_i\lambda_k \leqq \omega|\lambda|^2$; $\{(a_{011}(x, u), \cdots, a_{0mm}(x, u)) : u \in K\} = P$ *a.e. in* Ω;

(b) *assumptions* (b) *and* (c) *of Theorem* 3 *hold.*

Proof. From [1] we see that $\sup\{|z_n(u)(x_0, t_0)| : n = 0, 1, 2, \cdots, u \in U\} < \infty$. Therefore $m_n \equiv \inf G_n(U) \in R^1$ for every n. As in Theorem 3, for every $u \in U$ there exists a subsequence such that $z_n(u)(x, t) \to z_0(u)(x, t)$ uniformly on compact subsets of Q, and $z_n(u) \rightharpoonup z_0(u)$ in $L^2(H_0^1(\Omega))$: moreover, from (a), [10], [11] and [3], there exists $u_0 \in U$ such that (for some subsequence) $z_n(u_n)(x, t) \to z_0(u_0)(x, t)$ uniformly on compact subsets of Q, and $z_n(u) \rightharpoonup z_0(u)$ in $L^2(H_0^1(\Omega))$. Therefore, for every $u \in U$, $G_n(u) \to G_0(u)$ and $G_n(u_n) \to G_0(u_0)$ (for some subsequence). Since $m_n \leqq G_n(u_n) < \varepsilon_n + m_n$ and $m_n \leqq G_n(u)$ for every u and $n = 1, 2, \cdots$, we see that $\lim m_n = G_0(u_0) \leqq G_0(u)$ for every $u \in U$.

Acknowledgment. This paper was written during a stay at Brown University, Providence, Rhode Island, for which I wish to thank professors J. P. Cecconi and W. H. Fleming.

REFERENCES

[1] D. G. Aronson, *Non-negative solutions of linear parabolic equations*, Ann. Scuola Norm. Sup. Pisa, 22 (1968), pp. 607–694.

[2] A. G. Butkovskiy, *Distributed Control System*, American Elsevier, New York, 1969.

[3] C. Castaing, *Sur les multiapplications mesurables*, Rev. Internat. Recherche Opér., 1 (1967), pp. 91–126.

[4] M. Chicco, *Principio di massimo forte per sottosoluzioni di equazioni ellittiche di tipo variazionale*, Boll. Un. Mat. Ital., 22 (1967), pp. 368–372.

[5] ———, *Principio di massimo generalizzato e valutazione del primo autovalore, per problemi ellittici del secondo ordine di tipo variazionale*, Ann. Mat. Pura Appl. ser. IV, 87 (1970), pp. 1–10.

[6] W. H. Fleming, *Optimal control of partially observable diffusions*, this Journal, 6 (1968), pp. 194–214.

[7] A. Friedman, *Optimal control for parabolic equations*, J. Math. Anal. Appl., 18 (1967), pp. 479–491.

[8] O. A. Ladyzhenskaya, V. A. Solonnikov and N. N. Ural'ceva, *Linear and quasilinear equations of parabolic type*, Translations of Mathematical Monographs, American Mathematical Society, Providence, 1968.

[9] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[10] S. Spagnolo, *Sul limite delle soluzioni di problemi di Cauchy relativi all'equazione del calore*, Ann. Scuola Norm. Sup. Pisa, 21 (1967), pp. 657–699.

[11] ———, *Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche*, Ibid., 22 (1968), pp. 571–597.

[12] G. Stampacchia, *Équations elliptiques du second ordre à coefficients discontinus*, Sem. Math. Sup., Montréal, Quebec, 1966.

[13] T. Zolezzi, *Teoremi d'esistenza per problemi di controllo ottimo retti da equazioni ellittiche o paraboliche*, Rend. Sem. Mat. Univ. Padova, 44 (1970), pp. 155–173.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

FREDHOLM OPERATORS, ENCIRCLEMENTS, AND STABILITY CRITERIA*

JON H. DAVIS†

Abstract. This paper considers some applications of the theory of so-called Fredholm operators to the problem of input-output stability of linear feedback systems. It is shown that stability conditions may be obtained for certain classes of systems by calculating the *index* of an associated Fredholm operator. This index may be calculated by counting the encirclements of the origin by some curve in the complex plane.

Examples are given of applications to time-invariant problems with an unstable open loop, to the derivation of a version of the so-called converse of the circle criterion, and to a class of continuous time systems with a periodic linear feedback gain. In all three examples, stability (and instability) conditions are given in terms of encirclements (or lack of same) of a certain point by a locus in the complex plane.

1. General discussion. We consider the problem of input-output stability of feedback systems described by the functional equations

$$y = Ge,$$

$$e = u - Ky,$$

or, combining the above,

$$(I + KG)e = u. \tag{1}$$

It is known that under the hypothesis that the finite-time truncated version of the loop equation (1), written as

$$(I + KG)e_T = u_T, \qquad 0 \leqq t \leqq T, \tag{2}$$

has a unique solution for each truncation time T, the stability (or instability) of the system is determined completely by the invertibility of the operator $I + KG$ of (1). This invertibility condition may be equivalently stated either in terms of the causality of operators, or by considering (1) in the framework of Banach spaces defined on a half-axis in time. This latter point of view is the one adopted here, so that input-output stability of the system (1) is equivalent to the existence of a bounded inverse for the operator $I + KG$ (relative to its action on the half-axis Banach space). Formal definitions of the concept of input-output stability employed here are given in [1], [3], [5], and the specific result mentioned above is given in [5, Thm. 1]. Equivalent statements based on the notion of causality appear in [3, Thms. 4.1, 4.2].

The determination of the invertibility of an operator of the type that occurs in physical feedback systems is in general a very difficult mathematical problem, so that the equivalence of stability to the existence of an inverse can be considered more as an identification of the mathematical problem involved than as a general solution of the stability problem.

* Received by the editors February 16, 1971.

† Department of Mathematics, Queen's University, Kingston, Ontario, Canada.

There is one class of operators, however, for which invertibility is potentially much easier to check for than in the general case of an arbitrary operator defined in a Banach space. This is the class of Fredholm operators, also referred to as ϕ-operators in the Russian literature [6], [12].

DEFINITION. A closed linear operator T acting from a Banach space $\mathscr{X}$ to a Banach space $\mathscr{Y}$ is called a *Fredholm operator* if

(i) $\alpha(T)$, the dimension of $\eta(T)$, the null space of T, is finite;

(ii) $\beta(T)$, the codimension of $\mathscr{R}(T)$, the range of T, is finite.

The *index* of the Fredholm operator T is

$$\kappa = \alpha(T) - \beta(T).$$

Our interest in Fredholm operators stems from the fact that the invertibility of the Fredholm operator T may be checked easily (in fact, immediately) if the integers $\alpha(T)$ and $\beta(T)$ are known. In fact, in the case of the feedback stability problem, things are even easier than usual. If $I + KG$ is a Fredholm operator, then for systems such that the truncated loop equation (2) has a unique solution, $\alpha(I + KG)$ must be identically zero. This situation typically arises from the causality of KG, which results in (2) becoming an integral equation of Volterra type having a unique solution. In this case since $\alpha(I + KG) = 0$ we have $\kappa(I + KG) = -\beta(I + KG)$, and so $I + KG$ will be invertible if and only if $\kappa(I + KG) = 0$.

This reduction of the problem to that of simply calculating the index κ has a profound practical implication: while α and β are separately somewhat difficult to compute (see [8], for example), the integer κ may often be determined from an encirclement condition.

Fredholm operators were used in [4] to obtain certain results on systems where K and G represent bounded operators, and are implicit in [5]. Applications are made in the following section to both time-invariant and time-varying systems with an unstable open loop, and to continuous time systems with a periodic feedback gain.

There is some overlap between the specific results obtained in §§ 2.1 and 2.2 below, and those of other authors [1], [3], [21], although our method of derivation is new. The results in § 2.3 have not previously appeared.

2. Applications.

2.1. Time-invariant problems with unbounded open-loop operator. We consider in this section feedback systems described by the convolution equation

$$e(t) + k\int_0^t g(t - s)e(s)\,ds = u(t), \qquad t \geqq 0, \tag{3}$$

in the case that (3) represents a system made up of a cascade of a (generally open-loop unstable) finite-dimensional system and a convolution operator from the algebra $\alpha I \oplus L_1(0, \infty)$. (See Fig. 1.)

THEOREM 1. *Suppose that the Laplace transform of a function $g(\cdot)$ has the form $\hat{g}(s) = (q(s)/p(s))\hat{g}_1(s)$, where p and q are polynomials, degree q is less than degree p, p has no zero on the imaginary axis, and $\hat{g}_1(s)$ is the Laplace transform of an element*

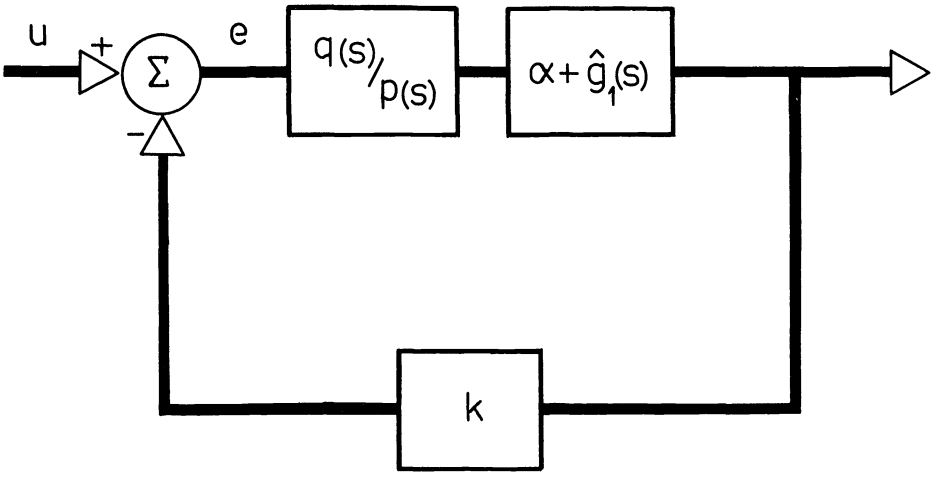


FIG. 1

of the convolution algebra $\alpha I \oplus L_1(0, \infty)$. *Then the feedback system described by the equation*

$$e(t) + k\int_0^t g(t-s)e(s)\,ds = u(t), \qquad t \geqq 0, \tag{3}$$

is input-output stable on $L_p(0, \infty)$, $1 \leqq p < \infty$, *if and only if the Nyquist criterion is satisfied for* $\hat{g}(s)$.

Comment. Since convolution by such a function $g(\cdot)$ in general represents an unbounded operator, a certain amount of care must be taken in defining the domain of the operator. The proof given below follows [12, Chap. IV, V and VI] in terminology, and freely uses results given there.

Proof of Theorem 1. The first task is to suitably define the domain of the convolution G. The appropriate choice of domain for G is the *range* of the so-called minimal closed extension of the differential operator $p(D)$ on $L_p(0, \infty)$. That this is the appropriate choice may be intuitively seen by recalling that formal Laplace transform manipulation in the equation

$$\left(1 + k\frac{q(s)}{p(s)}\hat{g}_1(s)\right)\hat{e}(s) = \hat{u}(s) \tag{4}$$

corresponds to differential equations with zero initial conditions. Further, the range of the minimal extension of $p(D)$ on $L_p(0, \infty)$, if $p(s)$ has no zero on the imaginary axis, is a closed subspace of $L_p(0, \infty)$ with codimension equal to the number of zeros of $p(s)$ in $\operatorname{Re}(s) > 0$. This follows since by Theorem VI. 5.6 of [12], the index of the maximal operator corresponding to $p(D)$ is equal to that of the minimal operator plus the degree of $p(s)$,

$$\kappa_{\max} = \kappa_{\min} + \deg p(s).$$

By Theorem VI.7.2, $\kappa_{\max}$ is just the number of zeros of $p(s)$ in Re $(s) < 0$. Since the minimal operator is 1-1 (Theorem VI.2.10), $\kappa_{\min}$ is just the negative of the codimension of the range of the minimal operator. Intuitively, the range of the minimal extension of $p(D)$ may be thought of as the set of functions whose Laplace transforms have right-half-plane zeros of sufficient multiplicity, so that enough cancellation of poles and zeros occurs in (4) to insure that (the input) $\hat{u}(s)$ is bounded in $L_p(0, \infty)$.

With this domain, $I + kG$ is a closed operator. Moreover, it is 1-1 (has no null space) on its domain, since Volterra integral equations of the second kind have a unique solution on a finite time interval if the kernel is an L_1 convolution kernel [14].

Consider now the case where $\hat{g}_1(s)$ has a rational Laplace transform, $\hat{g}_1(s) = q_1(s)/p_1(s)$. The minimal closed extensions of the operators $p(D)$ and $p(D)p_1(D)$ have the same range, since the polynomial $p_1(s)$ has no zero in the closed right-half-plane. Hence the range of $I + kG$ is identical to that of the minimal closed extension of the differential operator $\Delta(D) = p(D)p_1(D) + kq(D)q_1(D)$.

If $\Delta(s)$ has no zero on the imaginary axis, then the minimal closed extension of $\Delta(D)$ is a Fredholm operator, it is 1-1, and its index is given by the negative of the number of zeros of $\Delta(s)$ in Re $(s) > 0$ (see [12, Thms. VI.7.2 and VI.5.6]). If $\Delta(s)$ has a zero on the imaginary axis, then $\Delta(D)$ is not Fredholm.

From these observations and the general remarks of § 1, we conclude that $I + kG$ is a Fredholm operator with both indices zero if and only if the classical Nyquist condition is satisfied. This is so because the Nyquist condition simply counts zeros of $\Delta(s)$ in Re $(s) > 0$.

If the transform of $g_1(\cdot)$ is not rational, then we approximate g_1 in the L_1-norm by a function with rational transform (sufficiently closely so that the encirclements are the same). Since both "Fredholmness" and the index of a Fredholm operator are preserved under sufficiently small (relatively speaking, if the operator is unbounded) perturbations, we conclude that in the case where the Nyquist locus of $(q(s)/p(s))\hat{g}_1(s)$ does not intersect the point $-1/k$, the theorem is valid ([12, Theorem V.3.6] gives the required perturbation result).

If the Nyquist locus of $(q(s)/p(s))\hat{g}_1(s)$ intersects the point $-1/k$, then the argument given in [7, Theorem 12.1] may be easily modified to show that $I + kG$ is not a Fredholm operator, and $I + kG$ is not invertible.

Hence we conclude that the system (3) is input-output stable on $L_p(0, \infty)$ if and only if the Nyquist condition is satisfied for $\hat{g}(s) = (q(s)/p(s))\hat{g}_1(s)$.

2.2. Time-varying feedback with unbounded open-loop operator. In this section we consider the replacement of the constant feedback gain of the problem of the previous section with a time-varying one. Specifically, we consider the system

$$e(t) + k(t)\int_0^t g(t - s)e(s)\,ds = u(t), \qquad t \geqq 0, \tag{5}$$

with the same assumptions as in the previous section on the function $g(\cdot)$, and with the measurable function $k(\cdot)$ restricted by $0 < \alpha + \varepsilon \leqq k(t) \leqq \beta - \varepsilon$ (almost everywhere). This is essentially the "circle theorem" problem [1], [13], so that

it is not too surprising that there is considerable overlap in the results. What should be noted, however, is that Fredholm operator ideas provide a virtually immediate proof both of the circle theorem and its instability counterpart.

THEOREM 2 (The circle theorem and converse). *Suppose that a function* $g(\cdot)$ *satisfies the conditions of Theorem* 1 *above, and that the measurable function* $k(\cdot)$ *is restricted by* $0 < \alpha + \varepsilon \leqq k(t) \leqq \beta - \varepsilon$ *for some* α, β *and* $\varepsilon > 0$. *Let* ρ *denote the number of zeros of* $p(s)$ *in* $\operatorname{Re}(s) > 0$.

Then the feedback system described by the equation

$$e(t) + k(t)\int_0^t g(t-s)e(s)\,ds = u(t), \qquad t \geqq 0, \tag{5}$$

(i) *is input-output stable on* $L_2(0, \infty)$ *provided that the Nyquist locus of* $\hat{g}(s)$ *does not intersect the (closed) disc* $D[\alpha, \beta]$ *drawn through the points* $-1/\alpha$ *and* $-1/\beta$ *with center* $-\frac{1}{2}(1/\alpha + 1/\beta)$, *and encircles* $D[\alpha, \beta]$ ρ *times in the counterclockwise sense*;

(ii) *is input-output unstable if the Nyquist locus does not intersect* $D[\alpha, \beta]$ *and encircles it in the counterclockwise sense fewer than* ρ *times.*

Proof. The proof is a direct application of the basic perturbation result for the index of a Fredholm operator [12, Theorem V.3.6]. This states roughly that if the magnitude of the perturbation is bounded in terms of the magnitude of the original Fredholm operator, then the perturbed operator is Fredholm with the same index as the original.

Write the original system in the form

$$[I + \tfrac{1}{2}(\alpha + \beta)G + (K - \tfrac{1}{2}(\alpha + \beta)I)G]e = u,$$

and define the domain of the unbounded operator as in Theorem 1. By Theorem V.3.6 of [12], $I + KG$ is Fredholm with the same index as $I + \frac{1}{2}(\alpha + \beta)G$ provided that

$$\|(K - \tfrac{1}{2}(\alpha + \beta)I)Gx\| < \|(I + \tfrac{1}{2}(\alpha + \beta)G)x\|, \qquad x \in \mathscr{D}(G), \tag{6}$$

the domain of G. The above condition (6) simply requires that the Nyquist locus not intersect $D[\alpha, \beta]$, as is well known. Since the index of $I + \frac{1}{2}(\alpha + \beta)G$ is equal to the number of encirclements of the point $-2/(\alpha + \beta)$ minus the number of zeros of $p(s)$ in $\operatorname{Re}(s) > 0$, the result follows immediately.

Remark. The proof of Theorem 2 indicates the instability part of the circle criterion may be weakened to allow intersections of $D[\alpha, \beta]$ by the Nyquist locus. All that is required is that the point $-2/(\alpha + \beta)$ be encircled an "insufficient" number of times, and that (6) be satisfied. Various restrictions on $k(t)$ as in [15], [16], [17] should lead to instability counterparts of these theorems.

2.3. Systems with periodic feedback gain. We now restrict ourselves to feedback gains $k(t)$ which are periodic: $k(t + 1) = k(t)$ (again, almost everywhere). For a large class of open-loop stable systems of the form

$$e(t) + k(t)\int_0^t g(t-s)e(s)\,ds = u(t), \qquad t \geqq 0, \tag{5'}$$

with $k(t+1) = k(t)$, it is possible to give results for the time-varying system (5′) which are entirely analogous to those of § 2.1 for time-invariant systems. That is, it is possible to associate with (5′) a certain Fredholm operator, to give an explicit formula for the index of the Fredholm operator, and to prove that (5′) is stable if and only if it defines a Fredholm operator of zero index.

Since the mathematical details are somewhat involved, they have been largely reserved for the Appendix. The restrictions imposed and the main results will be stated in this section, and some implications will be discussed.

It turns out that the solution of the problem hinges on the use of a transform technique. However, the appropriate transform to use is neither a Fourier transform nor a z-transform, but rather something which is halfway between the two.

DEFINITION. For a function $g(\cdot)$ defined on R^1, we call $g(z;\tau) = \sum_{-\infty}^{\infty} z^{-n} g(n+\tau)$ the *H-transform* (for either "half" or "hybrid") of the function $g(\cdot)$.

PROPERTY 1. If $g(\cdot) \in L_1$, then $g(e^{i\theta};\cdot) \in L_1(0,1)$ for $0 \leqq \theta \leqq 2\pi$.

PROPERTY 2. If $g(\cdot) \in L_2$, then $g(e^{i\theta};\cdot) \in L_2(0,1)$ for almost all θ, $0 \leqq \theta \leqq 2\pi$.

Property 1 is immediate, and Property 2 requires Fubini's theorem, Plancherel's theorem, and the Riesz–Fischer theorem for proof.

It is easily seen that the Fourier coefficients of the function $g(e^{i\theta};\cdot)$ are given by $a_n = \hat{g}(i(2n\pi + \theta))$, where $\hat{g}(\cdot)$ is the Fourier transform of $g(\cdot) \in L_1$.

DEFINITION. We say that a function $g(\cdot) \in L_1$ defines an *F-kernel function* if

$$\int_0^1 \int_0^1 |g(e^{i\theta}; t-s)|^2 \, dt \, ds \leqq M < \infty$$

independent of θ.

The above condition insures that $g(e^{i\theta}; t-s) = k_\theta(t-s)$ defines a Hilbert–Schmidt kernel [9], [10]. This is equivalent to requiring that the series $\sum_{n=-\infty}^{\infty} |\hat{g}((2n\pi+\theta)i)|^2$ converge uniformly with respect to θ. This condition is stronger than requiring that $g(\cdot) \in L_1 \cap L_2$, in which case the series converges only almost everywhere. However, functions $g(\cdot) \in L_1$ with rational Fourier transforms belong to the class of F-kernel functions, so that any $g(\cdot) \in L_1$ may be approximated in the L_1-norm by a function defining an F-kernel.

Notation. If $k(\cdot) \in L_\infty(0,1)$ and $g(\cdot) \in L_1$ defines an F-kernel, then there is associated with the integral equation (in $L_2(0,1)$)

$$\alpha(t) - \lambda k(t) \int_0^1 g(e^{i\theta}; t-s)\alpha(s)\, ds = \beta(t), \qquad 0 \leqq t \leqq 1, \tag{7}$$

a so-called Fredholm determinant, which we denote by $\delta(\lambda; e^{i\theta})$ (see [9], [10]). (The terminology F-kernel arises from the fact that in this case the Fredholm determinant above is well-defined for each θ, $0 \leqq \theta \leqq 2\pi$.)

We are now in a position to give necessary and sufficient conditions for the stability of (5′). It is shown in the Appendix that the function $\delta(\lambda; z)$ plays a role with regard to the system (5′) exactly parallel to that of the function $1 + k\hat{g}(s)$ encountered in connection with the time-invariant systems of § 2.1. That is, if the locus of $\delta(\lambda; z)$ for $|z| = 1$ does not pass through the origin, then (5′) defines a Fredholm operator, whose index is given by the number of encirclements of the

origin made by the locus. Thus the result given below represents a generalization of the classical Nyquist encirclement condition to cover a system with a time-varying gain.

THEOREM 3. *Let* $k(\cdot)$ *be a function of period* $1, k(\cdot) \in L_\infty$, *and suppose that* $g(\cdot) \in L_1$ *defines an F-kernel. Then the feedback system*

$$e(t) + k(t)\int_0^t g(t-s)e(s)\,ds = u(t), \qquad t \geqq 0, \tag{5'}$$

is input-output stable on $L_2(0, \infty)$ *if and only if the following two conditions are satisfied*:

(i) $\delta(-1; e^{i\theta}) \neq 0, 0 \leqq \theta \leqq 2\pi$;

(ii) $\kappa = \dfrac{1}{2\pi}\displaystyle\int_0^{2\pi} d_\theta \arg \delta(-1; e^{i\theta}) = 0.$

Here $\delta(\lambda; z)$ *is the Fredholm determinant associated with the integral equation*

$$\alpha(t) - \lambda\int_0^1 k(t)g(z; t-s)\alpha(s)\,ds = \beta(t), \qquad 0 \leqq t \leqq 1.$$

Proof. According to Theorem A.3 of the Appendix to this paper, $I + KG$ for the system (5') has a bounded inverse on $L_2(0, \infty)$ if and only if (i) and (ii) above hold.

Remark 1. While the result of Theorem 3 has been stated in a form directly applicable to scalar input-scalar output systems, the result is valid also for vector input-vector output systems. The only modification necessary is in the definition of the F-kernel function class. In the vector input-vector output case the condition becomes

$$\int_0^1\int_0^1 \operatorname{tr}\,(G^*(e^{i\theta}; t-s)G(e^{i\theta}; t-s))\,dt\,ds \leqq M < \infty$$

independent of θ, $0 \leqq \theta \leqq 2\pi$, which becomes in the frequency domain the condition that the series

$$\sum_{n=-\infty}^{\infty} \operatorname{tr}\,[\hat{G}^*((2n\pi + \theta)i)\hat{G}((2n\pi + \theta)i)]$$

converge uniformly in θ, $0 \leqq \theta \leqq 2\pi$. The Fredholm determinant for the (vector) integral equation is then well-defined (in terms of operator traces, see [10]) and the proofs given in the Appendix carry through with virtually no modification.

Remark 2. Theorem 3 is a continuous time version of results previously obtained in a somewhat more elementary fashion for discrete systems [5].

Remark 3. A sufficient condition for stability may be obtained simply by requiring that the Hilbert–Schmidt norm of the kernel of the integral equation (7) be less than unity. That is, the condition that

$$\sup_{0 \leqq \theta \leqq 2\pi} \int_0^1\int_0^1 k^2(t)|g(e^{i\theta}; t-s)|^2\,dt\,ds < 1 \tag{8}$$

is sufficient for stability. This condition is comparable to the circle criterion in the case that the time-varying gain satisfies $|k(t)| < \beta$, so that the Nyquist locus must lie inside a circle of radius $1/\beta$ in order to guarantee stability. Since the Hilbert–Schmidt norm of the kernel $g(e^{i\theta}; t - s)$ is given by

$$\int_0^1 \int_0^1 |g(e^{i\theta}; t - s)|^2 \, dt \, ds = \sum_{-\infty}^{\infty} |\hat{g}((2n\pi + \theta)i)|^2,$$

condition (8) is capable of exploiting an averaging effect over the values assumed by the feedback gain. It is clear that (8) will be most useful for systems which show a large degree of high frequency attenuation, so that the quantity

$$\sup_{0 \leqq \theta \leqq 2\pi} \sum_{-\infty}^{\infty} |\hat{g}((2n\pi + \theta)i)|^2$$

is close to $\sup_\omega |\hat{g}(i\omega)|^2$.

(In interpreting the above remark, it should be borne in mind that the original problem has been implicitly time-scaled to make the feedback element of period 1. If the period is actually carried through the whole problem, then the "frequency interval" in the sums is just the frequency of the feedback element.)

It is of interest to note the "similar appearance" of condition (8), which in some sense involves an average of the square of the feedback gain in conjunction with the squared magnitude of the frequency response, and the condition given in [18] for stochastic systems, which involves the variance of the gain in conjunction with the squared magnitude of the frequency response.

Remark 4. In the case that $g(\cdot)$ has a rational transform, the finite-time integral equation (7) takes the form

$$\alpha(t) - \lambda k(t)\left[\left\{\int_0^1 \mathbf{c}'(\mathbf{I}z - e^{\mathbf{A}})^{-1} e^{\mathbf{A}(t-s+1)}\mathbf{b} + \int_0^t \mathbf{c}' e^{\mathbf{A}(t-s)}\mathbf{b}\right\}\alpha(s)\, ds\right] = \beta(t), \qquad 0 \leqq t \leqq 1. \tag{9}$$

If $k(t)$ is a constant, then the integral equation (9) has eigenvectors z^t with corresponding eigenvalue $\lambda_z = \mathbf{c}'(\mathbf{I} \log z - \mathbf{A})^{-1}\mathbf{b}$, so that the spectrum of the integral operator in (9) is just the Nyquist locus split into pieces of "length" equal to the feedback frequency. This is analogous to results obtained in the discrete analogue [5]. From this it is clear that Theorem 3 reduces to the Nyquist criterion result when the feedback gain is constant as it should.

3. Conclusions and suggestions for further research. This paper presents three examples of the application of the theory of Fredholm operators to feedback system stability problems, including both time-invariant and time-varying problems. It has been shown that this point of view leads naturally to stability criteria in terms of encirclement conditions analogous to the classical Nyquist criterion. It should be noted that for reasons of space not all of the Fredholm operator results currently available in the literature and directly applicable to feedback problems have been included in this presentation. For example, [20] contains encirclement-type conditions applicable to time-invariant problems involving a kernel containing delayed impulses. It would be of great interest to discover other

classes of time-varying feedback gain functions which lead to Fredholm operator problems. The class of almost periodic functions, for example, is of practical importance and is an obvious candidate to try.

One striking aspect of the results given here is the short proof and clear conceptual interpretation of the converse of the circle theorem possible via Fredholm operators. It is also clear that in order to apply these methods successfully to instability criteria it is necessary to analyze unbounded operators. This remark should be intuitively clear, since instability must involve unboundedness of the inverse of $I + KG$. As mentioned above, these methods should allow generalizations of the converse of the circle criterion analogous to the generalizations of the conditions already obtained [15], [16], [17] which are sufficient for stability.

The results given for systems with periodic feedback in terms of the Fredholm determinant function $\delta(\lambda; z)$ are completely analogous to plotting the Nyquist locus of $\det(\mathbf{I} + \mathbf{K}\hat{\mathbf{G}}(i\omega))$ for the vector input–output version of the time-invariant problem. The function $\delta(\lambda; z)$ is an entire function of λ with coefficients which are analytic functions of z (in the case of a Volterra-type kernel) so that it may be possible to solve the equation $\delta(\lambda; z) = 0$ to get $\lambda = h(z)$ for some analytic $h(\cdot)$. If this were possible, then the stability criterion would assume the form of checking for encirclements of λ by the Nyquist locus of $h(z)$. This form is analogous to checking encirclements of the point $-1/k$ in the time-invariant problem, and would yield information about the onset of instability with scaling of the feedback gain. If such an $h(z)$ could be defined (via an implicit function theorem), then there would remain questions of its effective computation, approximate calculation, error estimates, and so forth.

It would be interesting to extend the periodic system results to L_p-spaces for $p \neq 2$. The route of this extension is presently difficult to see, since the Hilbert–Schmidt operator theory which has been used is closely associated with Hilbert spaces.

Appendix. Spectrum of the product of a periodic multiplier and a convolution operator on $L_2(0, \infty)$. In this Appendix we consider the Fredholm theory associated with operators defined by

$$KGx(t) = k(t) \int_0^\infty g(t - s)x(s)\, ds, \qquad t \geqq 0, \tag{A.1}$$

where $k(t + 1) = k(t)$ is a bounded measurable function, and $g(\cdot) \in L_1(-\infty, \infty)$.

The results obtained are analogous to those obtained in [8] for systems of convolution equations on a half-axis, and the proofs follow the general outlines of those in [8].

For this reason proofs are abbreviated where possible, although points of difference will be explained. The main difficulty in extending the method of [8] to cover the present case lies in deciding what function of a complex variable is a proper analogue for the determinant function encountered in [8], since this function is what determines the index of the associated Fredholm operator. The idea that a certain Fredholm determinant is an appropriate choice comes from consideration of the discrete analogue of the problem [5].

Since the H-transform of a function defined in § 2.3 above is closely related both to the Fourier transform and the z-transform, it is to be expected that the H-transform of the convolution of two functions would be closely related to the H-transforms of the two functions involved.

LEMMA A.1. *If*

$$y(t) = \int_{-\infty}^{\infty} g(t - s)x(s)\,ds, \qquad x(\cdot) \in L_2, \quad g(\cdot) \in L_1,$$

then

$$y(z;\tau) = \int_0^1 g(z;\tau - s)x(z;s)\,ds;$$

that is, if $y = g * x$, *then* $y(z;\cdot) = g(z;\cdot) * x(z;\cdot)$.

This result is analogous to the usual result, with the exception that a finite time interval convolution replaces the usual pointwise multiplication of transforms.

The main usefulness of the H-transform for our purposes is that it commutes with multiplication by a periodic function of period 1. In fact, this is the motivation behind its definition.

LEMMA A.2. *If*

$$x(\cdot) \in L_2(-\infty, \infty),$$

and

$$k(t + 1) = k(t), \qquad k(\cdot) \in L_\infty,$$

then

$$(kx)(z;\tau) = k(\tau)x(z;\tau).$$

Applying the previous lemmas to the equation

$$x(t) - \lambda k(t) \int_0^\infty g(t - s)x(s)\,ds = y(t)$$

gives

$$x(z;\tau) - \lambda k(\tau)(g(z;\cdot) * x(z;\cdot))(\tau) = y(z;\tau). \tag{A.2}$$

Equation (A.2) has the form of an integral equation in $L_2(0, 1)$ with an additional parameter z. In order to make the problem tractable, (A.2) must be an equation which can be conveniently analyzed. In particular, we require that for each value of $z = e^{i\theta}, 0 \leqq \theta \leqq 2\pi$, equation (A.2) be an integral equation of Hilbert–Schmidt-type [9], [10], so that the original convolution kernel $g(\cdot) \in L_1$ must be an F-kernel, as defined in § 2.3 above.

DEFINITION. We let $\delta(\lambda; e^{i\theta})$ denote the so-called *Fredholm determinant* associated with the integral equation (A.2). (This function is identical to $\delta(\lambda)$ of [9], and differs by a substitution of λ^{-1} for λ in the function φ_λ of [10, p. 1036].)

LEMMA A.3. *The Fredholm determinant* $\delta(\lambda; e^{i\theta})$ *is a jointly continuous function of the two variables* λ *and* θ, *provided* $g(\cdot) \in L_1$ *defines an F-kernel.*

The above lemma follows readily from the continuity properties given in [9].

THEOREM A.1. *Suppose that* $k(\cdot) \in L_\infty$ *is periodic of period* 1, *and that* $g(\cdot) \in L_1$ *is an F-kernel. Then the operator* $(I - \lambda KG): L_2(0, \infty) \to L_2(0, \infty)$ *defined by*

$$(I - \lambda KG)x(t) = x(t) - \lambda k(t) \int_0^\infty g(t - s)x(s)\, ds, \qquad t \geqq 0,$$

is a Fredholm operator on $L_2(0, \infty)$ *provided that*

$$\delta(\lambda; e^{i\theta}) \neq 0, \qquad 0 \leqq \theta \leqq 2\pi,$$

where $\delta(\lambda; z)$ *is the Fredholm determinant for the integral equation*

$$\alpha(\tau) - \lambda k(\tau) \int_0^1 g(z; \tau - \sigma)\alpha(\sigma)\, d\sigma = \beta(\tau), \qquad 0 \leqq \tau \leqq 1,$$

in $L_2(0, 1)$.

Remark. Theorem A.1 above is the analogue of Theorem 2.1 of [8], and its proof differs mainly in that a generalization of Wiener's well-known result on invertible elements in an L_1-convolution algebra is required in the present case, while Wiener's result suffices in [8].

Proof of Theorem A.1. With the equation

$$x(t) - \lambda k(t) \int_0^\infty g(t - s)x(s)\, ds = y(t), \qquad t \geqq 0, \tag{A.3}$$

we associate the equation obtained by extending the above to the whole real axis:

$$\xi(t) - \lambda k(t) \int_{-\infty}^\infty g(t - s)\xi(s)\, ds = \eta(t), \qquad -\infty < t < \infty. \tag{A.4}$$

Taking the H-transform of the above gives

$$\xi(z; \tau) - \lambda \int_0^1 k(t)g(z; \tau - \sigma)\xi(z; \sigma)\, d\sigma = \eta(z: \tau), \qquad 0 \leqq \tau \leqq 1. \tag{A.5}$$

The operator $(I - \lambda kg(z; \cdot) *)$ on the left side of (A.5) may be imbedded in the ring of operators of the form $\sum_{-\infty}^{\infty} z^{-n} A_n$, with A_n operators such that $\sum_{-\infty}^{\infty} \|A_n\| < \infty$ (see [11]). Here, A_n is the operator defined on $L_2(0, 1)$ by

$$A_n x(t) = k(t) \int_0^1 g(n + t - \sigma)x(\sigma)\, d\sigma, \qquad 0 \leqq t \leqq 1,$$

and

$$\|A_n\| \leqq \|k\|_\infty \cdot \|g_n\|_{L_1(-1,1)},$$

so that

$$\sum_{-\infty}^{\infty} \|A_n\| \leqq \|k\|_\infty \cdot \|g\|_{L_1} \cdot 2.$$

According to the theorem of Bochner and Phillips [11, Thm. 1], the operator $(I - \lambda kg(z; \cdot) *)$ has an inverse in the ring if and only if the operator $(I - \lambda kg(e^{i\theta}; \cdot) *)$ has an inverse (bounded) for each fixed θ, $0 \leqq \theta \leqq 2\pi$. This is analogous to the result of Wiener that shows that the reciprocal of an absolutely

convergent Fourier series is an absolutely convergent Fourier series if and only if the original series does not vanish for any fixed value of its argument.

By the hypothesis that $g(\cdot) \in L_1$ is an F-kernel, the operator $(I - \lambda k g(e^{i\theta}; \cdot) *)$ for each fixed value of θ reduces to an integral operator on $L_2(0, 1)$ with a Hilbert–Schmidt kernel. Hence $(I - \lambda k g(e^{i\theta}; \cdot) *)$ has an inverse for each fixed value of θ if and only if the Fredholm determinant of the integral equation does not vanish, i.e., $\delta(\lambda; e^{i\theta}) \neq 0$. Therefore, $(I - \lambda k g(z; \cdot) *)$ has an inverse in the ring (of power series with absolutely convergent operator coefficients encountered above). This means that the operator $I - \lambda KG$ defined by (A.4) has a bounded inverse on $L_2(-\infty, \infty)$ provided that $\delta(\lambda; e^{i\theta}) \neq 0$, $0 \leqq \theta \leqq 2\pi$.

From this point, the argument is identical to that in [8]. Rewrite (A.4) as a system of two equations, one of which contains (A.3). These equations are identical to those on the bottom of p. 228 in [8], except for a multiplication of the integrals by the periodic function $k(\cdot)$. Again, this system of two equations is equivalent to (A.4), and the "off-diagonal" operators are compact, which is obvious, since the product of a bounded and a compact operator is compact. If $\delta(\lambda; e^{i\theta}) \neq 0$, then $I - \lambda KG$ is invertible on $L_2(-\infty, \infty)$, and hence is a Fredholm operator of zero index. Under this condition, we conclude that the original operator defined by (A.3) is a Fredholm operator using the same argument as [8].

Now that it has been established that if $\delta(\lambda; e^{i\theta}) \neq 0$, the operator $I - \lambda KG$ is a Fredholm operator, it remains to calculate the index of the operator. Since the function $\delta(\lambda; e^{i\theta})$ plays a role in the present investigation analogous to the function $\det(\mathbf{I} - \lambda \mathbf{K}\hat{\mathbf{G}}(i\omega))$ in [8], it is to be expected that the index of the Fredholm operator $I - \lambda KG$ can be found by means of $\delta(\lambda; z)$.

THEOREM A.2. *If $\delta(\lambda; e^{i\theta}) \neq 0$, $0 \leqq \theta \leqq 2\pi$, then the index of the Fredholm operator $I - \lambda KG$:*

$$x(t) \to x(t) - \lambda k(t) \int_0^\infty g(t - s)x(s)\, ds, \qquad t \geqq 0,$$

acting on $L_2(0, \infty)$ is given by

$$\kappa = \frac{1}{2\pi} \int_0^{2\pi} d_\theta \arg \delta(\lambda; e^{i\theta}),$$

that is, by the number of encirclements of the origin by the function $\delta(\lambda; z)$.

Proof. By general results [6], [8], the index of a Fredholm operator is invariant under perturbations of sufficiently small norm. That is, for each bounded Fredholm operator U, there exists a $\delta > 0$ such that if $\|U - V\| < \delta$, then V is also a Fredholm operator, and $\kappa(V) = \kappa(U)$. We approximate our Fredholm operator by one whose index is known by the following method. Let $g_M(\cdot) \in L_1$ denote the function obtained from $g(\cdot) \in L_1$ by replacing the values assumed by $g(\cdot)$ by zero for $|t| > M$. Pick M so large that $\|g_M - g\|_{L_1} < \delta/4$.

Next approximate the Hilbert–Schmidt kernel $k(\tau)g_M(e^{i\theta}; \tau - s)$ in the Hilbert–Schmidt norm by a degenerate kernel $KG_N(e^{i\theta}; t - s)$ constructed with respect to the orthonormal system $\{e^{2\pi i n t}\}_{n=-N}^{N}$, and choose N so large that, if $\delta_N(\lambda; e^{i\theta})$ is the Fredholm determinant for the degenerate approximating kernel,

the following two conditions hold:

(i) index $\delta(\lambda; e^{i\theta}) = \text{index } \delta_N(\lambda; e^{i\theta})$;

(ii) $\|\lambda KG(e^{i\theta}) - \lambda(KG)_N(e^{i\theta})\|_{\text{HS}} < \delta/4$.

That this is possible follows from [9] and the continuity properties of the Fredholm determinant.

The degenerate kernel integral equation

$$x(e^{i\theta}; \tau) - \lambda \int_0^1 (KG)_N(e^{i\theta}; \tau - s)x(e^{i\theta}; s)\, ds = y(e^{i\theta}; \tau)$$

is equivalent to a discrete convolution problem on a half-line. Specifically, it corresponds to the propagation of the first $2N + 1$ Fourier coefficients in the development of the unknown function $\xi(\cdot)$ in the equation

$$\xi(t) - \lambda k(t) \int_0^\infty g_M(t - s)\xi(s)\, ds = \eta(t)$$

over each successive time interval of unit length.

If we denote the z-transform of the corresponding $(2N + 1) \times (2N + 1)$-dimensional matrix sequence by $(\mathbf{K\hat{G}})_N(z)$, we note that $(\mathbf{K\hat{G}})_N(z)$ is a power series of which the highest degree terms are $z^{\pm M}$. This is so because of the truncation of g_M.

By [8], we see that the index of the discrete convolution operator $I - \lambda(KG)_N$ is given by

$$\kappa_N = \text{index } \{\det (\mathbf{I} - \lambda(\mathbf{K\hat{G}})_N(e^{i\theta}))\}.$$

However, the Fredholm determinant corresponding to the degenerate kernel is closely related to the above determinant. It follows from [9] that

$$\det (\mathbf{I} - \lambda(\mathbf{K\hat{G}})_N(e^{i\theta})) = e^{-\text{tr}(\mathbf{K\hat{G}})_N(e^{i\theta})}\delta_N(\lambda; e^{i\theta}). \tag{A.6}$$

Since $(\mathbf{K\hat{G}})_N(z)$ contains terms of order only up to $z^{\pm M}$, the expression

$$\text{tr } (\mathbf{K\hat{G}})_N(e^{i\theta})$$

is an absolutely convergent Fourier series. Therefore,

$$\text{tr } (\mathbf{KG})_N(e^{i\theta}) = f_1(e^{i\theta}) + f_2(e^{i\theta}),$$

where $f_1(e^{i\theta})$ admits a holomorphic extension $f_1(z)$ in $|z| < 1$, continuous on $|z| \leqq 1$, and $f_2(e^{i\theta})$ admits a holomorphic extension $f_2(z)$ in $|z| > 1$, continuous on $|z| \geqq 1$. Hence,

$$e^{-\text{tr}(\mathbf{K\hat{G}})_N(e^{i\theta})} = e^{-f_1(e^{i\theta})}\, e^{-f_2(e^{i\theta})}$$

is a product of two functions, one of which is the boundary value of an entire function of z, and the other of which is an entire function of $1/z$. Moreover, the first function does not vanish in $|z| \leqq 1$, and the second does not vanish in $|z| \geqq 1$. Since the index of a function which is a boundary value of an analytic function counts the difference between the number of poles and zeros inside the contour traversed, we conclude from (A.6) that

$$\text{index det } (\mathbf{I} - \lambda(\mathbf{K\hat{G}})_N(e^{i\theta})) = \text{index } \delta_N(\lambda; e^{i\theta}). \tag{A.7}$$

Equation (A.7) shows that the index of the operator $I - \lambda KG$ is given by the index of $\delta(\lambda; e^{i\theta})$, since the approximating operator was chosen so that

$$\text{index } \delta_N = \text{index } \delta$$

and index $(I - \lambda KG) = \text{index } (I - \lambda(KG)_N)$.

LEMMA A.4. *If*

$$\delta(\lambda; e^{i\theta_0}) = 0, \qquad 0 \leqq \theta_0 \leqq 2\pi,$$

then $I - \lambda KG$ is not a Fredholm operator on $L_2(0, \infty)$.

Proof. The corresponding proof of [7, Thm. 12.1] carries over directly.

We remark that Theorems A.1 and A.2 involving the Fredholm character and index of the equation

$$x(t) - \lambda k(t) \int_0^\infty g(t - s)x(s)\, ds = y(t), \qquad t \geqq 0,$$

hold regardless of whether the function $g(\cdot)$ vanishes for negative values of its argument or not. As mentioned in § 1 above, if a Fredholm operator is 1-1, then its invertibility can be determined simply from its index. In case $g(x) \equiv 0$ for $x < 0$, then the operator

$$KG: x(t) \to k(t) \int_0^t g(t - s)x(s)\, ds,$$

defined on $L_2(0, \tau)$ for any finite τ, is quasi-nilpotent. This follows directly from results given in [14], and shows immediately that $I - \lambda KG$ is 1-1. Combining these remarks with the results above gives the following.

THEOREM A.3. *Let $g(\cdot) \in L_1$ be an F-kernel, and let $k(\cdot) \in L_\infty$ satisfy $k(t + 1) = k(t)$ (almost everywhere). Then the operator*

$$(I - \lambda KG): x(t) \to x(t) - \lambda k(t) \int_0^t g(t - s)x(s), \qquad t \geqq 0,$$

defined on $L_2(0, \infty)$ has a bounded inverse if and only if the following two conditions are satisfied:

(i) $\delta(\lambda; e^{i\theta}) \neq 0, \qquad 0 \leqq \theta \leqq 2\pi,$

(ii) $\kappa = \dfrac{1}{2\pi} \displaystyle\int_0^{2\pi} d_\theta \arg (\delta(\lambda; e^{i\theta})) = 0.$

Here $\delta(\lambda; e^{i\theta})$ is the Fredholm determinant defined above.

Remark. It should be clear from the proofs above that the same results hold for vector (finite-dimensional) input–vector output systems. All that is necessary is to interpret the Fredholm determinant as that appropriate for such a system (see [10], for example), and to further complicate the notation involved in the approximation argument of Theorem A.2.

REFERENCES

[1] G. ZAMES, *On the input–output stability of non-linear time varying feedback systems, Part I*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238.

[2] G. ZAMES AND P. L. FALB, *Stability conditions for systems with monotone and slope restricted nonlinearities*, this Journal, 6 (1968), pp. 89–108.

[3] J. C. WILLEMS, *Stability, instability, invertibility and causality*, this Journal, 7 (1969), pp. 645–671.

[4] J. H. DAVIS, *Stability of linear feedback systems*, Doctoral thesis, Dept. of Mathematics, MIT, Cambridge, Mass., 1970.

[5] ———, *Stability conditions derived from spectral theory; Discrete systems with periodic feedback*, this Journal, 10 (1972), pp. 1–13.

[6] I. C. GOHBERG AND M. G. KREIN, *The basic propositions on defect numbers, root numbers and indices of linear operators*, Amer. Math. Soc. Transl. (2), 13 (1960), pp. 185–264.

[7] M. G. KREIN, *Integral equations on a half-line with kernel depending on the difference of the arguments*, Ibid., 22 (1962), pp. 163–288.

[8] I. C. GOHBERG AND M. G. KREIN, *Systems of integral equations on a half-line with kernels depending on the difference of the arguments*, Ibid., 14 (1960), pp. 217–287.

[9] F. SMITHIES, *The Fredholm theory of integral equations*, Duke Math J., 8 (1941), pp. 107–130.

[10] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part II*, Interscience, New York, 1963.

[11] S. BOCHNER AND R. S. PHILLIPS, *Absolutely convergent Fourier expansions for non-commutative normed rings*, Ann. of Math., 43 (1942), pp. 409–418.

[12] S. GOLDBERG, *Unbounded Operators*, McGraw-Hill, New York, 1966.

[13] R. W. BROCKETT AND H. B. LEE, *Frequency domain instability criteria for time varying and nonlinear systems*, Proc. IEEE, 55 (1967), pp. 604–619.

[14] J. R. RINGROSE, *Compact linear operators of the Volterra type*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 44–55.

[15] G. ZAMES AND M. FREEDMAN, *Logarithmic variation criteria for the stability of systems with time varying gains*, this Journal, 6 (1968), pp. 487–507.

[16] J. C. WILLEMS, *On the asymptotic stability of the null solution of linear differential equations with periodic coefficients*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 65–72.

[17] M. GRUBER AND J. L. WILLEMS, *On a generalization of the circle criterion*, Proc. Fourth Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1966, pp. 827–835.

[18] J. C. WILLEMS AND G. BLANKENSHIP, *Frequency domain stability criteria for stochastic systems*, submitted for publication.

[19] H. BATEMAN, *A formula for the solving function of a certain integral equation of the second kind*, Messenger of Math., 37 (1908), pp. 179–186.

[20] I. C. GOHBERG AND I. A. FELDMAN, *On Wiener Hopf integral-difference equations*, Soviet Math. Dokl., 9 (1968), no. 6, pp. 1312–1316.

[21] C. A. DESOER AND M. Y. WU, *Stability of multiple-loop feedback linear time-invariant systems*, J. Math. Anal. Appl., 23 (1968), pp. 121–129.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

EXISTENCE THEOREMS FOR MULTIDIMENSIONAL CONTROL SYSTEMS WITH LOWER-DIMENSIONAL CONTROLS*

RICHARD F. BAUM†

Abstract. Existence theorems are proved for multidimensional Lagrange control systems, in which the state is a function of several independent variables in a fixed domain, while in contrast, the control is a function of only some of the independent variables (that is, u is a lower-dimensional control). As usual in Lagrange problems, the cost functional is a multiple integral, the state equation is a system of partial differential equations, with assigned boundary conditions, and constraints may be imposed on the values of the state and control variables.

1. Introduction. In this paper, we prove existence theorems for multidimensional Lagrange control systems, in which the unknown, or state variable, is a vector-valued function $x(t) = (x^1, \cdots, x^n)(t)$, $t = (t^1, \cdots, t^v)$, of $v \geqq 1$ independent variables, $t \in G$, the domain. In contrast, we assume that the control variable, u, is a function of only some of the independent variables, say $(t^1, \cdots, t^d) = r$, $d < v$, so that the control function is of the form $u(t) = u(r) = (u^1, \cdots, u^m)(r)$ (that is, u is a lower-dimensional control). As usual, the pair $x(t), u(r)$ satisfies certain state equations—here, a system of partial differential equations, in normal form, certain smoothness and measurability properties, certain constraints on the values that x and u may assume, and certain boundary conditions. The cost functional has the usual Lagrange form $I[x, u] = \int_G f_0(t, x, u)\, dt$. We are concerned with the problem of finding the minimum of $I[x, u]$ in suitable classes of admissible pairs $x(t), u(r)$.

In applications, such control systems may arise if we are unable to ascertain the values of certain independent variables t^j, or due to lack of technology, we are unable to act upon this information. The former situation arises naturally in search problems, and in § 6, Example 3, we consider a search problem of this type with a moving target. Such systems may also arise in stochastic multidimensional systems; Example 4, in § 6, concerns this situation.

In § 2 we introduce notations, and give an introductory discussion of the above control systems; in § 4 we consider these control systems in complete detail. In § 3 we prove a closure theorem for these systems and establish a number of corollaries; these results are in turn used in § 5 to derive an existence theorem, as well as several corollaries. Finally, in § 6, we consider several examples in which we apply the results of our existence theorems. Necessary conditions for these control systems have been developed in [2].

Butkovsky, in [3, § 1.2], also considers necessary conditions for systems with lower-dimensional controls; these problems are somewhat different from the ones we examine here.

2. Description of the system. In this section, we discuss the structure of a class of multidimensional systems in which the domain of the control is of lower

* Received by the editors January 8, 1971.

† Department of Industrial Engineering, University of Michigan, Ann Arbor, Michigan 48104. This work was sponsored by the United States Air Force under Grants AF-AFOSR-69-1767-A and AFOSR-69-1662.

dimensionality than the domain of the state variable. This discussion is introductory; the full set of details for these systems is given in §§ 4 and 5.

Let the domain G be a bounded open subset of the t-space E_v. We require that the boundary ∂G of G satisfy certain smoothness conditions, in particular, that G be of class K_l [10, Chap. 1]. Let us suppose that for some reason, such as lack of information, the control can be a function of only some of the t variables, say for convenience, $t^1, t^2, \cdots, t^d$, $1 \leqq d < v$. Because of this, we decompose $t = (t^1, \cdots, t^v)$ into two parts, $t = (r, s)$, $r = (t^1, \cdots, t^d) \in E_d$, $s = (t^{d+1}, \cdots, t^v) \in E_e$, $d + e = v$, and we set $R = \{r : (r, s) \in G \text{ for some } s \in E_e\}$, $S = \{s : (r, s) \in G \text{ for some } r \in E_d\}$.

For each $t \in \operatorname{cl} G$ = closure of G, let $A(t)$ be a nonempty subset of the x-space E_n, with $x = (x^1, \cdots, x^n)$, the space variable. The set $A = \{(t, x) : t \in \operatorname{cl} G, x \in A(t)\} \subset E_{v+n}$ is then the constraint set. For each r in cl R, let the control set $U(t) = U(r)$, $t = (r, s)$, be a nonempty subset of the u-space E_m, $u = (u^1, \cdots, u^m)$, the control variable. Let $M = \{(t, x, u) : (t, x) \in A, u \in U(t) = U(r)\}$. For notational convenience, if $H \subset E_l$, and $(h^1, \cdots, h^l) \in H$, then for any $c < l$, let $H_{(h^1, \cdots, h^c)}$ be the section of H determined by $(h^1, \cdots, h^c)$, that is, let $H_{(h^1, \cdots, h^c)} = \{(h^{c+1}, \cdots, h^l) : (h^1, \cdots, h^c, h^{c+1}, \cdots, h^l) \in H\} \subset E_{l-c}$. Similar notation holds for any subset of indices.

We wish to consider systems of partial differential equations satisfied by certain vector functions $x(t) = (x^1(t), \cdots, x^n(t))$, $u(r) = (u^1(r), \cdots, u^m(r))$. In particular, we wish to consider $u = u(t) = u(r)$ as a function of r alone, that is, we ask, for any $r \in R$, that $u(r, s_1) = u(r, s_2)$ for all $s_1, s_2 \in G_r$. To this end, assume that for each $i = 1, \cdots, n$, we are given a finite collection $\{\alpha\}_i$ of indices, $\alpha = (\alpha_1, \cdots, \alpha_v)$, $1 \leqq |\alpha_i| \leqq l_i \leqq l$, with $|\alpha| = \alpha_1 + \cdots + \alpha_v$. Let the total number of indices $\{\alpha\}_i$, $i = 1, \cdots, n$, be N. Let $f(t, x, u) = (f_{i\alpha})$, the state function, be a continuous vector function from M into E_N, and let $f_0(t, x, u)$, the cost function, be a continuous function from M into E_1. Also suppose we are given a set (B) of boundary conditions on ∂G involving the boundary values of $x^i(t)$ and certain partial derivatives of $x^i(t)$, $i = 1, \cdots, n$, which we shall specify below (see § 4 for full details).

We shall be interested in pairs of functions $x(t) = (x^1(t), \cdots, x^n(t))$, $t \in G$ (trajectories), $u(r) = (u^1(r), \cdots, u^m(r))$, $r \in R$ (controls), which satisfy certain smoothness conditions, constraints, boundary conditions, and systems of partial differential equations. In particular, we shall call a pair of functions $x(t), u(r)$, $t = (r, s) \in G$, an admissible pair if: (i) each component $x^i(t)$, $t \in G$, belongs to a Sobolev space $W^{l_i}_{p_i}(G)$, $1 \leqq l_i \leqq l, p_i > 1, i = 1, \cdots, n$; in particular, each component $x^i(t)$ of $x(t)$ is L_{p_i}-integrable in G and possesses the generalized partial derivatives $D^\alpha x^i(t)$ of all orders $\alpha \in \{\alpha\}_i$, all L_{p_i}-integrable in G for certain $p_i > 1$, $i = 1, \cdots, n$ (see [8], [10, Chap. 1]); (ii) $u(r)$ is measurable in R (a function of r alone); (iii) the space constraint $(t, x(t)) \in A$ is satisfied a.e. in G; (iv) the control constraint $u(r) \in U(r)$ is satisfied a.e. in R; (v) the state equation (a system of partial differential equations in normal form) $D^\alpha x^i(t) = f_{i\alpha}(t, x(t), u(r))$ is satisfied a.e. in G, $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n$; (vi) $f_0(t, x(t), u(r))$ is L_1-integrable in G. We shall consider classes Ω of admissible pairs for which (vii) $x^i(t)$ and $D^\alpha x^i(t)$, $1 \leqq |\alpha| \leqq l_i \leqq l$, satisfy the set (B) of boundary conditions on ∂G. For each $(x, u) \in \Omega$, we shall associate the cost functional

$$I[x, u] = \int_G f_0(t, x(t), u(r))\, dt.$$

We are concerned with the problem of the minimum of $I[x, u]$ in Ω, that is, with the determination of pairs $x(t), u(r)$ in Ω so that $I[x, u] \leqq I[\tilde{x}, \tilde{u}]$ for all $(\tilde{x}, \tilde{u}) \in \Omega$.

3. Closure theorems. In order to include the behavior of the cost function f_0 directly into our system, we shall here augment the control system described in § 2. Specifically, we let $x = (x^1, \cdots, x^n) = (y, z)$, where $y = (x^1, \cdots, x^\sigma)$, $z = (x^{\sigma+1}, \cdots, x^n)$. We may think of the y-vector as corresponding to the state vector x of § 2, and of the z-vector as corresponding to the behavior of the cost functional. Accordingly, we assume that the state function $f(t, y, u)$ depends only on $x^1, \cdots, x^\sigma$, and for a given trajectory $x(t) = (y(t), z(t))$, the vector $y(t)$ possesses generalized partial derivatives $D^\alpha y^i(t)$ as described in § 2. We impose a different set of assumptions on z, the remaining $n - \sigma$ components of x.

We shall need certain properties of set functions. Given any set F in a linear space E, we shall denote by cl F and co F the closure of F and the convex hull of F, respectively.

For every $(t_0, x_0) \in A$, and $\delta > 0$, let $N_\delta(t_0, x_0)$ denote the closed δ-neighborhood of radius δ in A, that is, the set of all $(t, x) \in A$ at a distance $\leqq \delta$ from (t_0, x_0).

Let $F(t, x)$ be a variable set in a Euclidean space E, a set function of (t, x) in A. We shall say that F is an upper semicontinuous function of (t, x) at the point $(t_0, x_0) \in A$ if, given $\varepsilon > 0$, there is a $\delta = \delta(t_0, x_0, \varepsilon) > 0$ such that $(t, x) \in N_\delta(t_0, x_0)$ implies $F(t, x) \subset [F(t_0, x_0)]_\varepsilon$, where $[F]_\varepsilon$ denotes the closed ε-neighborhood of F in E. For $X \subset E_N$, let $F(t, X) = \{z : z = f(t, x) \text{ for some } x \in X\} \subset E$.

Again, let $F(t, x), (t, x) \in A$, be a variable set in E. For every $\delta > 0$, let $F(t_0, x_0; \delta) = \cup F(t, x)$, where the union is taken for all $(t, x) \in N_\delta(t_0, x_0)$. We shall say that $F(t, x)$ has property (U) at $(t_0, x_0) \in A$ if

$$F(t_0, x_0) = \bigcap_{\delta > 0} \text{cl } F(t_0, x_0; \delta).$$

We shall say that $F(t, x)$ has property (Q) at $(t_0, x_0) \in A$ if

$$F(t_0, x_0) = \bigcap_{\delta > 0} \text{clco } F(t_0, x_0; \delta).$$

We shall say that $F(t, x)$ has property (U), (Q) in A if $F(t, x)$ has property (U), (Q) at every (t, x) in A, respectively. A set $F(t, x)$ satisfying property (U) is necessarily closed, and a set satisfying property (Q) is necessarily closed and convex.

CLOSURE THEOREM. *Let the domain G be a bounded open set of the t-space E_v of some class K_l, $l \geqq 1$, with $t = (r, s)$, $r \in E_d$, $s \in E_e$, $d + e = v$. Let $R = \{r : (r, s) \in G \text{ for some } s \in E_e\}$, $S = \{s : (r, s) \in G \text{ for some } r \in E_d\}$, and $\hat{S} = \{\hat{s} \in S : G_{\hat{s}} = R\} \subset S$. Assume* meas $\hat{S} > 0$. *Let $A(t)$ be a nonempty closed subset of the y-space E_σ, $t \in$ cl G, and assume that the constraint set $A = \{(t, y) : t \in \text{cl } G, y \in A(t)\} \subset E_v \times E_\sigma$ is also closed. Let $\tilde{A} = A \times E_{n-\sigma}$, and let the control set $U(t) = U(r), t = (r, s)$ be a nonempty closed subset of the u-space E_m defined for every $r \in$ cl R. Assume that $U(r)$ satisfies property* (U) *in* cl R. *Let $M = \{(t, y, u) : (t, y) \in A, u \in U(r) = U(t)\}$. Let $\{\alpha\}_i$ be a finite collection of indices, $i = 1, \cdots, n$, $\alpha = (\alpha_1, \cdots, \alpha_v)$, $1 \leqq |\alpha_i| \leqq l_i$ with $l_i \leqq l$, defined for every $i = 1, \cdots, n$, and let $\tilde{N}$ be the total number of elements in $\{\alpha\}_i$, $i = 1, \cdots, n$.*

For every $i = s + 1, \cdots, n$, $\{\alpha\}_i$ consists of only one element $\alpha_0 = (1, \cdots, 1)$. Hence, if N is the total number of elements $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n$, then

$\tilde{N} = N + (n - \sigma)$. *Let* $\{\gamma\}_i$ *be those indices from* $\{\alpha\}_i$ *with* $\alpha_{d+1} = \alpha_{d+2} = \cdots = \alpha_v = 0$, $i = 1, \cdots, \sigma$, *and let* K *be the total number of elements in* $\{\gamma\}_i$, $i = 1, \cdots, \sigma$, *and assume* $K > 0$. *Let* $f(t, y, u) = (f_{i\alpha}), (\alpha \in \{\alpha\}_i, i = 1, \cdots, n)$, *be a continuous* N*-vector function on the set* M, *and assume that the set* $\tilde{\tilde{Q}}(t, y) = \{z = (z^1, \cdots, z^{\tilde{N}}) : (z^1, \cdots, z^N) = f(t, y, u),\ z^i \geqq f_{i\alpha_0}(t, y, u),\ i = N + 1, \cdots, \tilde{N},$ *for some* $u \in U(t) = U(r)\}$ *is a closed convex subset of* $E_{\tilde{N}}$ *for every* $(t, y) \in A$, *and that* $\tilde{\tilde{Q}}(t, y)$ *satisfies property* (Q) *in* A. *Assume that* $f_{i\alpha}(t, y, u) \geqq -c_{i\alpha}$, $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, \sigma$, *for all* $(t, y, u) \in M$ *and some constants* $c_{i\alpha} \geqq 0$. *Assume further that*: (Γ) *whenever* $(t, y, u_1), (t, y, u_2)$ *are in* M, *and* $f_{i\gamma}(t, y, u_1) = f_{i\gamma}(t, y, u_2)$ *for all* $\gamma \in \{\gamma\}_i$, $i = 1, \cdots, \sigma$, *then* $f_{i\alpha}(t, y, u_1) = f_{i\alpha}(t, y, u_2)$ *for all* $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n$ (see also Remark 3.2).

Also assume that: (Δ) *for* $\gamma \in \{\gamma\}_i$, $i = 1, \cdots, \sigma$,

$$f_{i\gamma}(t, y, u) = g_{i\gamma}(t, y) + w_{i\gamma}(s, y) k_{i\gamma}(r, u), \tag{3.1}$$

or in vector notation,

$$f_\gamma(t, y, u) = g_\gamma(t, y) + w_\gamma(s, y) \otimes k_\gamma(r, s),$$

where $w_i(s, y) \neq 0$ *for all* $(s, y) \in \{(s, y) : (r, s, y) \in A$ *for some* $r \in R\}$, *and for* $a, b \in E_k$, $a \otimes b$ *denotes* $(a^1 b^1, a^2 b^2, \cdots, a^k b^k) \in E_k$. *Let* $g(t, y) = (g_{i\gamma})$, $w(s, y) = (w_{i\gamma})$, $k(r, u) = (k_{i\gamma}), \hat{f}(t, y, u) = (f_{i\gamma}), \gamma \in \{\gamma\}_i, i = 1, \cdots, \sigma$. *For any fixed* $s_1, s_2 \in S$, *and* $(r, s_1, y_1), (r, s_2, y_2) \in A$, *let* $Q(r, y_1, y_2; s_1, s_2) = \{z = (z^1, \cdots, z^{2K}) : (z^1, \cdots, z^K) = \hat{f}(r, s_1, y_1, u), (z^{K+1}, \cdots, z^{2K}) = \hat{f}(r, s_2, y_2, u)$ *for some* $u \in U(r)\} \subset E_{2K}$. *We assume that*: ($\Xi$) *for any* $s_1, s_2 \in S$, $Q(r, y_1, y_2; s_1, s_2)$ *satisfies property* (Q) *for all* $(r, y_1, y_2) \in H(s_1, s_2) = \{(r, y_1, y_2) : (r, s_i, y_i) \in A, i = 1, 2\}$ (see also Corollaries 3.1 and 3.2).

Let $x_k(t) = (y_k(t), z_k(t)), u_k(r), t = (r, s) \in G, k = 1, 2, \cdots$, *be a sequence of pairs of vector functions with* $x_k(t) = (x_k^1, \cdots, x_k^n)(t) \in E_n$, $y_k(t) = (x_k^1, \cdots, x_k^\sigma)(t)$, $z_k(t) = (x_k^{\sigma+1}, \cdots, x_k^n)(t)$, $u_k(r) = (u_k^1, \cdots, u_k^m)(r) \in E_m$, *and let* $x_k(t) \in L_1(G)$, *and* $u_k(r)$ *be measurable in* R. *Assume that each component* $x_k^i(t)$ *possesses generalized partial derivatives* $D^\alpha x_k^i(t)$ *for* $\alpha \in \{\alpha\}_i$, $t \in G$, $D^\alpha x_k(t) \in L_1(G)$, *such that, for a.a.* $t = (r, s) \in G$,

$$(t, y_k(t)) \in A, \qquad u_k(r) \in U(r), \tag{3.2}$$

$$D^\alpha x_k^i(t) = f_{i\alpha}(t, y_k(t), u_k(r)), \tag{3.3}$$

for all $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n, k = 1, 2, \cdots$.

Let $x(t) = (x^1, \cdots, x^n)(t), t \in G$, *be a vector function whose components* $x^i(t) \in L_1(G)$ *possess generalized partial derivatives* $D^\alpha x^i(t), \alpha \in \{\alpha\}_i, i = 1, \cdots, n$, *and assume that*

$$x_k^i(t) \to x^i(t) \quad \textit{pointwise a.e. in } G \textit{ as } \quad k \to \infty, \qquad i = 1, \cdots, \sigma, \tag{3.4}$$

$$D^\alpha x_k^i(t) \to D^\alpha x^i(t) \quad \textit{weakly in } \quad L_1(G), \quad \alpha \in \{\alpha\}_i, \quad \textit{as } k \to \infty, \quad i = 1, \cdots, \sigma. \tag{3.5}$$

Moreover, assume there is a countably dense collection $[t]$ *in* E_1 *so that for all points* $\{t\}, t \in G$, *of the form* $t = (t^1, \cdots, t^v), t^j \in [t], j = 1, \cdots, v$, *we have*

$$x_k^i(t) \to x^i(t) \quad \textit{at every} \quad t \in \{t\} \quad \textit{as} \quad k \to \infty, \quad i = \sigma + 1, \cdots, n, \tag{3.6}$$

and for all intervals $\{I\}$, $I \subset G$, *with vertices in* $\{t\}$, *we have*

$$\int_I D^\alpha x_k^i(t)\,dt \to \int_I D^\alpha x^i(t)\,dt \quad \text{as } k \to \infty \tag{3.7}$$

with $\alpha = \alpha_0 \in \{\alpha\}_i$, $i = \sigma + 1, \cdots, n$.

Assume that there is a decomposition $x^i(t) = Z^i(t) + S^i(t)$, $i = \sigma + 1, \cdots, n$, x^i *and* S^i *both of class* $L_1(G)$, $Z^i(t)$ *possessing generalized partial derivative* $D^{\alpha_0} Z^i$ *of class* $L_1(G)$, *and* $S^i(t)$ *singular.*

Then there exists a measurable vector function $u(r) = (u^1, \cdots, u^m)(r)$, $r \in R$, *such that, a.e. for* $t = (r, s) \in G$,

$$(t, y(t)) \in A, \quad u(r) \in U(r), \tag{3.8}$$

$$D^\alpha x^i(t) = f_{i\alpha}(t, y(t), u(r)), \quad a \in \{\alpha\}_i, \qquad i = 1, \cdots, \sigma, \tag{3.9}$$

$$D^\alpha Z^i(t) \geqq f_{i\alpha}(t, y(t), u(r)), \qquad \alpha = \alpha_0, \quad i = \sigma + 1, \cdots, n. \tag{3.10}$$

Proof. From the hypotheses on $x_k(t)$ and $x(t)$, $t \in G$, it follows from Closure Theorem 2 of Cesari [6], that there exists a measurable vector function $u(t)$ such that, for a.a. $t = (r, s) \in G$,

$$(t, y(t)) \in A, \quad u(t) \in U(t) = U(r), \tag{3.11}$$

$$D^\alpha x^i(t) = f_{i\alpha}(t, y(t), u(t)), \qquad i = 1, \cdots, \sigma, \tag{3.12}$$

$$D^\alpha Z^i(t) \geqq f_{i\alpha}(t, y(t), u(t)), \qquad \alpha = \alpha_0, \quad i = \sigma + 1, \cdots, n. \tag{3.13}$$

It remains to show that $u(t)$ may be taken as a function of r alone. To do this, we shall show that for some fixed $\bar{s} \in \hat{S}$, $u(\cdot, \bar{s})$ generates the trajectories $y(\cdot, s)$, $Z(\cdot, s)$ in the sense that for a.a. $s \in S$,

$$D^\alpha x^i(r, s) = f_{i\alpha}(r, s, y(r, s), u(r, \bar{s})), \qquad i = 1, \cdots, \sigma, \tag{3.14}$$

$$D^{\alpha_0} Z^i(r, s) \geqq f_{i\alpha_0}(r, s, y(r, s), u(r, \bar{s})), \qquad i = \sigma + 1, \cdots, n, \tag{3.15}$$

a.e. in R, and moreover, (3.14) and (3.15) hold for a.a. $(r, s) \in G$. The theorem then follows by choosing $u(r)$ as $u(r, \bar{s})$, $r \in R$.

Let us now show that for a.a. $s' \in S$, $u(\cdot, \bar{s})$ generates $x(\cdot, s')$, $Z(\cdot, s')$ for an appropriate $\bar{s} \in S$. In particular, let S_0 be that subset of S such that $\hat{s} \in S_0$ implies $G_{\hat{s}}$ is of class K_l and $D^\alpha x^i(r, \hat{s}) = f_{i\alpha}(r, \hat{s}, y(r, \hat{s}), u(r, \hat{s}))$, $D^{\alpha_0} Z^i(r, \hat{s}) \geqq f_{i\alpha_0}(r, \hat{s}, y(r, \hat{s}), u(r, \hat{s}))$ a.e. in $G_{\hat{s}}$. Then meas S_0 = meas S. Let $s' \in S_0$, $\bar{s} \in S_0 \cap \hat{S}$, where $S_0 \cap \hat{S} \neq \varnothing$ since meas $\hat{S} > 0$.

We shall now show that $u(\cdot, \bar{s})$ generates $y(\cdot, s')$, $Z(\cdot, s')$. To this end, let us consider the control system with constraint set $\bar{A} = \operatorname{cl} G_{s'} \times A(r, \bar{s}) \times A(r, s')$, a closed subset of the rY-space $E_d \times E_{2\sigma}$, with $r \in E_d$, and $Y = (Y^1, \cdots, Y^\sigma, Y^{\sigma+1}, \cdots, Y^{2\sigma})$, the space variable, in $E_{2\sigma}$. For each $r \in \operatorname{cl} G_{s'}$, let the control set $U(r)$ be the same as before, with $u = (u^1, \cdots, u^m)$, the control variable. Let $\bar{M} = \{(r, Y, u) : (r, Y) \in \bar{A}, u \in U(r)\}$. Let $F(r, Y, u) = (F_{i\delta}), (\delta \in \{\delta\}_i, i = 1, \cdots, 2\sigma)$ be the continuous $2K$-vector function on the set $\bar{M}$ defined by

$$F(r, Y, u) = (\hat{f}(r, \bar{s}, Y^1, \cdots, Y^\sigma, u), \hat{f}(r, s', Y^{\sigma+1}, \cdots, Y^{2\sigma}, u)),$$

$$\hat{f}(t, y, u) = (f_{i\gamma}), \quad \{\delta\}_i = \{\delta\}_{i+n} = \{\gamma\}_i, \qquad \gamma \in \{\gamma\}_i, \quad i = 1, \cdots, \sigma.$$

We wish to consider the class Π of pairs $Y(r), u(r), r \in G_{s'}$, satisfying (a) $Y^i(r) \in L_1(G_{s'})$ and possesses generalized partial derivatives $D^\delta Y^i(t) \in L_1(G_{s'})$, $\delta \in \{\delta\}_i$, $i = 1, \cdots, 2\sigma$, (b) $u(r)$ is measurable in $G_{s'}$, (c) $(r, Y(r)) \in \bar{A}$ for a.a. $r \in G_{s'}$, (d) $u(r) \in U(r)$ for a.a. $r \in G_{s'}$, (e) the state equation $D^\delta Y^i(r) = F_{i\delta}(r, Y(r), u(r))$ is satisfied a.a. in $G_{s'}$ for $\delta \in \{\delta\}_i$, $i = 1, \cdots, 2\sigma$. By the construction of this control system, it is clear that $(Y_k(r), u_k(r)) = (y_k(r, \bar{s}), y_k(r, s'), u_k(r)) \in \Pi$, $r \in G_{s'} = G_{s'} \cap G_{\bar{s}}$.

We wish to show that $Y(r) = (y(r, \bar{s}), y(r, s'))$ can be generated by some measurable $u(r)$, that is, there is a $u(r)$ such that $(Y(r), u(r)) \in \Pi$. Since $A(t)$ is closed, $t \in G$, it follows that $\bar{A}$ is closed. By relation (3.4) of the hypotheses, $Y_k(j) \to Y(r)$ pointwise a.e. in $G_{s'}$, and by relation (3.5), $D^\delta Y_k^i(r) \to D^\delta Y^i(r)$ weakly in $L_1(G)$ as $k \to \infty$, $i = 1, \cdots, 2\sigma$. By assumption, $Q(r, Y) = Q(r, (Y^1, \cdots, Y^\sigma), (Y^{\sigma+1}, \cdots, Y^{2\sigma}); \bar{s}, s') = F(r, Y, U(r))$ satisfies property (Q) in $G_{s'}$. Hence, we can apply Closure Theorem 1 of Cesari [6] to conclude that there exists a measurable control function $u(r)$ so that

$$D^\delta Y^i(r) = F_{i\delta}(r, Y(r), u(r)), \quad u(r) \in U(r), \tag{3.16}$$

a.e. in $G_{s'}$, $\delta \in \{\delta\}_i$, $i = 1, \cdots, 2\sigma$. By the construction of $Y(r)$, $F(r, Y, u)$ and (3.16), it follows that $\hat{f}(r, \bar{s}, y(r, \bar{s}), u(r, \bar{s})) = \hat{f}(r, \bar{s}, y(r, \bar{s}), u(r))$, $\hat{f}(r, s', y(r, s'), u(r, s')) = \hat{f}(r, s', y(r, s'), u(r))$ a.e. in $G_{s'}$. Hence

$$k(r, u(r, \bar{s})) = k(r, u(r)) = k(r, u(r, s')) \tag{3.17}$$

a.e. in $G_{s'}$, and therefore $\hat{f}(r, s', y(r, s'), u(r, s')) = \hat{f}(r, s', y(r, s'), u(r, \bar{s}))$ a.e. in $G_{s'}$

Since s' was chosen arbitrarily in S_0, it follows from (3.17) that

$$k(r, u(r, \bar{s})) = k(r, u(r, s)) \tag{3.18}$$

for all $s \in S_0$, $r \in G_s$. Let H be the subset of G in which relation (3.18) is satisfied. Since $u(r, \bar{s})$, $u(r, s)$ are measurable in G, and since for each $s \in \{s : (r, s) \in H \text{ for some } r \in R\}$, H_s is of full measure, it follows that H is measurable, with meas H = meas G. Thus, relation (3.18) holds for a.a. $(r, s) \in G$. Hence,

$$\hat{f}(r, s, y(r, s), u(r, s)) = \hat{f}(r, s, y(r, s), u(r, \bar{s}))$$

a.e. in G. By assumption (Γ), this then implies that

$$f(r, s, y(r, s), u(r, s)) = f(r, s, y(r, s), u(r, \bar{s})) \tag{3.19}$$

a.e. in G. Thus, for a.a. $(r, s) \in G$, relations (3.14) and (3.15) follow from relations (3.12), (3.13) and (3.19). By our previous remarks the closure theorem is thereby proved.

Remark 3.1. The proof of the closure theorem remains essentially the same if we replace the assumption that $\tilde{\tilde{Q}}(t, y)$ satisfies property (Q) in A with the assumption that $\tilde{\tilde{Q}}(t, y)$ is an upper semicontinuous function of (t, y) in A (see [5]). Similarly, the theorem remains valid if we replace assumption (Ξ) with the assumption that for any fixed $s_1, s_2 \in S$, $Q(r, y_1, y_2; s_1, s_2)$ is a closed convex subset of E_{2K} for every $(r, y_1, y_2) \in H(s_1, s_2)$, and that $Q(r_1, y_1, y_2; s_1, s_2)$ is an upper semicontinuous function of (r, y_1, y_2) in $H(s_1, s_2)$ (see [5, Closure Theorem 1]). We shall use this remark to eliminate assumption (Ξ).

COROLLARY 3.1 (for compact control sets). *Let the assumptions of the closure theorem be in effect, with the following additions*: (i) *let* $A(t)$, $t \in G$, *and* A *be compact sets*; (ii) *let the control set* $U(r)$ *be an upper semicontinuous function of* r *in* R;

(iii) *omit assumption* (Ξ), *and assume instead only that* $k(r, U(r))$ *is a convex subset of* E_n *for each* $r \in \text{cl } R$. *Then the conclusions of the closure theorem remain valid.*

Proof. By the hypotheses on $U(r)$, $U(r)$ satisfies property (U) in R, and M is a compact set [4, § 4], [5, § 2]. Hence $f_{i\alpha_0}, i = \sigma + 1, \cdots, n$, are bounded on M, say, $|f_{i\alpha_0}(t, y, u)| \leqq M_0$. Thus, the proof of the closure theorem remains unchanged if we replace $\tilde{\tilde{Q}}(t, y)$ with $\tilde{\tilde{Q}}(t, y; M_0) = \{z = (z^1, \cdots, z^N) : (z^1, \cdots, z^N) = f(t, y, u), M_0 \geqq z^i \geqq f_{i\alpha_0}(t, y, u), i = N + 1, \cdots, \tilde{N}$, for some $u \in U(r)\}$. Clearly, $Q(t, y; M_0)$ remains a closed convex subset of E_N for each $(t, y) \in A$. Hence, $Q(t, y; M_0)$ is then an upper semicontinuous function of (t, y) in A [5, § 4]. Moreover, $k(r, U(r))$, and hence, $Q(r, u_1, y_2; s_1, s_2)$, is a closed subset for each $(r, y_1, y_2) \in H(s_1, s_2)$, and by (iii), by taking $t = (r, s_1)$ and $t = (r, s_2)$, it follows that $Q(r, y_1, y_2; s_1, s_2)$ is also convex.

For consider any $(r, y_1, y_2) \in H(s_1, s_2)$ and any $u_1, u_2 \in U(r)$. We have for any $0 \leqq \theta \leqq 1$,

$$
\begin{aligned}
&\theta F(r, (y_1, y_2), u_1) + (1 - \theta) F(r, (y_1, y_2), u_2) \\
&\quad = \theta(\hat{f}(r, s_1, y_1, u_1), \hat{f}(r, s_2, y_2, u_1)) + (1 - \theta)(\hat{f}(r, s_1, y_1, u_2), \hat{f}(r, s_2, y_2, u_2)) \\
&\quad = (g(r, s_1, y_1) + w(s_1, y_1) \otimes (\theta k(r, u_1) + (1 - \theta) k(r, u_2)), g(r, s_2, y_2) \\
&\quad + w(s_2, y_2) \otimes (\theta k(r, u_1) + (1 - \theta) k(r, u_2))) \\
&\quad = (g(r, s_1, y_1) + w(s_1, y_1) \otimes k(r, u_0), g(r, s_2, y_2) + w(s_2, y_2) \otimes k(r, u_0)) \\
&\quad = (\hat{f}(r, s_1, y_1, u_0), \hat{f}(r, s_2, y_2, u_0)) \\
&\quad = F(r, (y_1, y_2), u_0),
\end{aligned}
$$

where, by the assumption that $k(r, U(r))$ is convex, there exists a $u_0 \in U(r)$ such that $k(r, u_0) = \theta k(r, u_1) + (1 - \theta) k(r, u_2)$. Hence $Q(r, y_1, y_2; s_1, s_2)$ is also an upper semicontinuous function of (r, y_1, y_2) in $H(s_1, s_2)$ [5, § 2]. By Remark 3.1, the corollary is thereby proved.

COROLLARY 3.2. *Let the assumptions of the closure theorem be in effect, with the exception that assumption* (Ξ) *be replaced by the assumptions*: (i) $k(r, U(r))$ *is a bounded upper semicontinuous function of* r *in* R, (ii) $k(r, U(r))$ *is a convex subset of* E_N *for each* $r \in R$. *Then the conclusions of the closure theorem remain valid.*

Proof. As shown in the proof of Corollary 3.1, $Q(r, y_1, y_2; s_1, s_2)$ is convex for each $r \in R$ by assumption (ii). By Remark 3.1, our proof is completed once we show that $Q(r, y_1, y_2; s_1, s_2)$ is an upper semicontinuous function of (r, y_1, y_2) in $H(s_1, s_2)$. To this end, consider $r^0, y_1^0, y_2^0, s_1, s_2$, with $(r^0, y_1^0, y_2^0) \in H(s_1, s_2)$. We wish to show that for $\varepsilon > 0$, there is a $\delta = \delta(r^0, y_1^0, y_2^0) = \delta(r^0, y_1^0, y_2^0; s_1, s_2) > 0$, such that for $|(r, y_1, y_2) - (r^0, y_1^0, y_2^0)| < \delta$, $Q(r, y_1, y_2; s_1, s_2) \subset [Q(r_1^0, y_1^0, y_2^0; s_1, s_2)]_\varepsilon$. For this purpose, let $\varepsilon > 0$ be given. By the continuity of $g(t, y)$ and $w(s, y)$, we can choose $\delta_1 = \delta_1(r^0, y_1^0, y_2^0, \varepsilon) > 0$ so that

$$
\begin{aligned}
&|g(r, s_i, y_i) - g(r^0, s_i, y_i^0)| < \varepsilon_1 = \varepsilon/6, && i = 1, 2, \\
&|w(s_i, y_i) - w(s_i, y_i^0)| < \varepsilon_2 = \varepsilon/6nV_0, && i = 1, 2,
\end{aligned} \tag{3.20}
$$

where $|k(r, u)| \leqq V_0 < \infty, r \in R, u \in U(r)$. From the upper semicontinuity of $k(r, U(r))$, we can choose $\delta_2 = \delta_2(r^0, \varepsilon) > 0$ so that for $|r - r^0| < \delta_2$,

$$
k(r, U(r)) \subset [k(r^0, U(r^0)]_{\varepsilon_3}, \tag{3.21}
$$

$\varepsilon_3 = \varepsilon/3n[|w(s_1, y_1^0)| + |w(s_2, y_2^0)|]$. Let $\delta = \min(\delta_1, \delta_2)$ and let $|(r, y_1, y_2) - (r^0, y_1^0, y_0^2)| < \delta$. Let $q \in Q(r, y_1, y_2; s_1, s_2)$. Then $q = (\hat{f}(r, s_1, y_1, u), \hat{f}(r, s_2, y_2, u))$ for some $u \in U(r)$. By (3.21), there is a $u^0 \in U(r^0)$ such that $|k(r, u) - k(r^0, u^0)| < \varepsilon_3$. Hence if $q^0 = (\hat{f}(r^0, s_1, y_1^0, u^0), \hat{f}(r^0, s_2, y_2^0, u^0))$, then

$$
\begin{aligned}
|q - q^0| &\leqq |g(r, s_1, y_1) - g(r^0, s_1, y_1^0)| + |g(r, s_2, y_2) - g(r^0, s_2, y_2^0)| \\
&\quad + |w(s_1, y_1) \otimes k(r, u) - w(s_1, y_1^0) \otimes k(r^0, u^0)| + |w(s_2, y_2) \otimes k(r, u) \\
&\qquad - w(s_2, y_2^0) \otimes k(r^0, u^0)| \\
&\leqq \varepsilon_1 + n|k(r, u)|\,|w(s_1, y_1) - w(s_1, y_1^0)| + n|w(s_1, y_1^0)|\,|k(r, u) \\
&\quad - k(r^0, u^0)| + n|k(r, u)|\,|w(s_2, y_2) - w(s, y_2^0)| \\
&\quad + n|w(s_2, y_2^0)|\,|k(r, u) - k(r^0, y^0)| \\
&\leqq 2\varepsilon_1 + \varepsilon_2 n 2V_0 + \varepsilon_3 n[|w(s_1, y_1^0)| + |w(s_2, y_2^0)|] < \varepsilon.
\end{aligned} \tag{3.22}
$$

Hence, $Q(r, y_1, y_2; s_1, s_2)$ is an upper semicontinuous function of (r, y_1, y_2) in $H(s_1, s_2)$. By previous remarks, Corollary 3.2 is thereby proved.

Remark 3.2. (a) The proof of the closure theorem remains essentially the same if we replace the explicit form of $f_{i\gamma}(t, y, u)$, $\gamma \in \{\gamma\}_i$, $i = 1, \cdots, \sigma$, in (3.1) with the assumption that for any $\gamma \in \{\gamma\}_i$, $i = 1, \cdots, \sigma$, whenever $(r, s_j, y_j, u_j) \in M$, $j = 1, 2$, and there is a $\bar{u} \in U(r)$ so that

$$
\begin{aligned}
f_{i\gamma}(r, s_1, y_1, u_1) &= f_{i\gamma}(r, s_1, y_1, \bar{u}), \\
f_{i\gamma}(r, s_2, y_2, u_2) &= f_{i\gamma}(r, s_2, y_2, \bar{u}),
\end{aligned}
$$

then it follows that

$$
f_{i\gamma}(r, s_1, y_1, u_1) = f_{i\gamma}(r, s_1, y_1, u_2).
$$

In particular, the assumption that $w(s, y) \neq 0$ for all $s \in S$, $y \in A_s$, can be weakened to $w(s, y) \neq 0$ for all $s \in \hat{S}$, $y \in A_s$.

(b) If $f_{i\alpha}(t, y, u)$, $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n$, has the same form as $f_{i\gamma}(t, y, u)$ given in (3.1), $\gamma \in \{\gamma\}_i$, $i = 1, \cdots, \sigma$, then assumption (Γ) reduces to the requirement that if $k_{i\gamma}(r, u_1) = k_{i\gamma}(r, u_2)$ for all $\gamma \in \{\gamma\}_i$, $i = 1, \cdots, \sigma$, then $k_{i\alpha}(r, u_1) = k_{i\alpha}(r, u_2)$ for all $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n$.

Remark 3.3. (a) The closure theorem, with its corollaries, may be extended to domain G for which meas $\hat{S} = 0$, but which is made up of sets G_k with meas $\hat{S}_k > 0$, and satisfying the property (F): that for a.a. $r \in R$, the hyperplane G_r intersects at most one component G_k. In particular, G may consist of components $G_1, \cdots, G_\gamma$, this collection satisfying property (F), with meas $\hat{S}_k > 0$, $k = 1, \cdots, \gamma$, and in each of these G_k, there may be a different system of $\{\alpha\}_i$, $i = 1, \cdots, \sigma$, and functions $f_{i\alpha}$.

(b) We may weaken the assumptions on G and $f_{i\alpha}$ in another way. Let G_k, $k = 1, \cdots, \gamma$, be finitely many open bounded subsets of E_v, this collection satisfying property (F). For each G_k, suppose meas $\hat{S}_k > 0$, and that there is a given set $\{\alpha\}_{ik}$, $i = 1, \cdots, n$, $k = 1, \cdots, \gamma$, and a system $f_{i\alpha}$ of functions. Let $\{F_\tau\}$ be the set of all possible nonempty intersections of the form $F = G_{k_1} \cap G_{k_2} \cap \cdots \cap G_{k_p}$, $1 \leqq p \leqq \gamma$. Each F_τ is a nonempty bounded open subset of E_v, and $\{F_\tau\}$ has a finite number of elements. Moreover, by property (F), each $F \in \{F_\tau\}$ is of measure zero. Thus, in each G_k, there is a different system of partial differential equations, and

these systems are compatible. Moreover, $f_{i\alpha}$ can now be assumed to be only sectionally continuous on each G_k, coinciding on each set F_τ with functions which are continuous on cl F_τ (see [6, Remark 3.4]). This discussion also extends to countable collections $\{G_k\}$ of open sets, with $\cup \partial G_k$ of measure zero.

(c) Let us suppose G consists of the union of components G_k, $k = 1, \cdots, \gamma$, as described above in (a) or (b), except that now we do not require that property (F) be satisfied. (If $\{G_k\}$ is given as in (b), we must now ask that the union of the various differential systems over $\{F_\tau\}$ be compatible.) Furthermore, suppose that except for only one G_k, say G_1, the corresponding state functions $(f_{ik\alpha})$, $\alpha \in \{\alpha\}_{ik}$, $i = 1, \cdots, n$, $k = 2, \cdots, \gamma$, do not depend upon u, that is, $f_{ik\alpha} = f_{ik\alpha}(t, x)$. Then the closure theorem and its corollaries extend to these systems (without requiring that property (F) be satisfied), since for each $r \in R$, we may extend by constancy the control $u(r)$ obtained in G_1 to all of G_r; this will have no effect on the other components G_k, $k \neq 1$, since the functions $(f_{ik\alpha})$, $k \neq 1$, do not contain u.

4. Admissible pairs. We shall continue to use the same notation of §§ 2 and 3. In addition to the state function $f(t, x, u) = (f_{i\alpha})$, we also consider a cost function $f_0(t, x, u)$, continuous on M, and let $\tilde{f}(t, x, u) = (f_0, f)$. We ask that each component $x^i(t)$ of $x(t)$, $i = 1, \cdots, n$, belong to the Sobolev class $W^{l_i}_{p_i}(G)$ for some given l_i and p_i, $1 \leqq l_i \leqq l, p_i > 1$, $i = 1, \cdots, n$. Consequently, each $x^i(t)$ and $D^\alpha x^i(t)$, $\alpha = (\alpha_1, \cdots, \alpha_v), 0 \leqq |\alpha| \leqq l_i - 1$, has boundary values Φ^i_α defined almost everywhere on the boundary ∂G of G, and each Φ^i_α is of class $L_{p_i}(\partial G)$.

Let (B) be a set of boundary conditions involving the boundary values of $x^i(t)$ and $D^\alpha x^i(t), 0 \leqq |\alpha| \leqq l_i - 1$. We require that the boundary conditions (B) satisfy a closure and an inequality property. The closure property is given by (P_1): if $x(t)$, $x_k(t), t \in G$, are n-vector functions, with $x^i(t), x^i_k(t) \in W^{l_i}_{p_i}(G)$, if $D^\beta x^i_k(t) \to D^\beta x^i(t)$ as $k \to \infty$ weakly in $L_{p_i}(G)$ for every β with $|\beta| = l_i$, and strongly in $L_{p_i}(G)$ for every β with $0 \leqq |\beta| \leqq l_i - 1$, and if the boundary values $\Phi^i_{k\alpha}$ of $x^i_k(t)$, $k = 1, \cdots, n$, $0 \leqq |\alpha| \leqq l_i - 1$, on ∂G satisfy the boundary conditions (B), then the boundary values Φ^i_α of $x^i(t), i = 1, \cdots, n, 0 \leqq |\alpha| \leqq l_i - 1$, also satisfy the boundary conditions (B).

The inequality property is given by (P_2): if $x(t), t \in G$, is any n-vector function satisfying boundary conditions (B), with components $x^i(t) \in W^{l_i}_{p_i}(G), p_i > 1$, $1 \leqq l_i \leqq l$, satisfying

$$\int_G |D^\beta x^i(t)|^{p_i} dt \leqq M_{i\beta}$$

for all $\beta = (\beta_1, \cdots, \beta_v)$ with $|\beta| = l_i, i = 1, \cdots, n$, and constants $M_{i\beta}$, then there are constants $M_{i\alpha}$ such that

$$\int_G |D^\alpha x^i(t)|^{p_i} dt \leqq M_{i\alpha}$$

for all $\alpha = (\alpha_1, \cdots, \alpha_v)$ with $0 \leqq |\alpha| \leqq l_i - 1, i = 1, \cdots, n$, where the constants $M_{i\alpha}$ depend only on p_i, v, all $M_{i\beta}, G$, and boundary conditions (B), but not on the function $x(t)$.

For example, if the boundary conditions (B) are defined by requiring that all derivatives $D^\alpha x^i(t)$ equal preassigned continuous boundary value functions Φ^i_α on ∂G, $\alpha = (\alpha_1, \cdots, \alpha_v), 0 \leqq |\alpha| \leqq l_i - 1, i = 1, \cdots, v$, then conditions (P_1) and (P_2) are satisfied.

A pair of functions $x(t) = (x^1(t), \cdots, x^n(t))$, $u(r) = (u^1(r), \cdots, u^m(r))$, $t = (r, s) \in G$, with $x^i(t) \in W^{l_i}_{p_i}(G)$, $u^j(r)$ measurable in R, satisfying $(t, x(t)) \in A$, $u(r) \in U(r)$, $D^\alpha x^i(t) = f_{i\alpha}(t, x(t), u(r))$, $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n$, a.e. in G, and $f_0(t, x(t), u(r)) \in L_1(G)$, is said to be admissible. A class Ω of admissible pairs, satisfying a set of boundary conditions (B), is said to be complete if, for any sequence $x_k(t), u_k(r)$, $k = 1, 2, \cdots$, of pairs from Ω and any admissible pair (x, u) such that $x_k \to x$ in the manner described in (P_1), then the pair (x, u) belongs to Ω. For example, the class of all admissible pairs is complete.

5. Existence theorems.

EXISTENCE THEOREM. *Let the domain G be a bounded open set of the t-space E_v of some class K_l, $l > 1$, with $t = (r, s)$, $r \in E_d$, $s \in E_e$, $d + e = v$. Let $R = \{r : (r, s) \in G$ for some $s \in E_e\}$, $S = \{s : (r, s) \in G$ for some $r \in E_d\}$, and $\hat{S} = \{\bar{s} \in S : G_{\bar{s}} = R\} \subset S$. Assume* meas $\hat{S} > 0$ (see also Remark 3.3). *Let $A(t)$ be a nonempty closed subset of the x-space E_n, $t \in$ cl G, and assume that the constraint set $A = \{(t, x) : t \in$ cl G, $x \in A(t)\} \subset E_v \times E_n$ is also closed.*

Let the control set $U(t) = U(r)$, $t = (r, s)$, be a nonempty closed subset of the u-space E_m defined for every $t \in$ cl R. Assume that $U(r)$ satisfies property (U) *in* cl R. *Let $M = \{(t, x, u) : (t, x) \in A, u \in U(r)\}$. Let $\{\alpha\}_i$ be a finite collection of indices, $i = 1, \cdots, n$, $\alpha = (\alpha_1, \cdots, \alpha_v)$, $0 \leqq |\alpha| \leqq l_i \leqq l$, and let N be the total number of elements in $\{a\}_i$, $i = 1, \cdots, n$. Let $\{\gamma\}_i$ be those indices from $\{\alpha\}_i$ with $\alpha_{d+1} = \alpha_{d+2} = \cdots = \alpha_v = 0$, $i = 1, \cdots, n$, let K be the total number of elements in $\{\gamma\}_i$, $i = 1, \cdots, n$, and assume $K > 0$.*

Let $\tilde{f}(t, x, u) = (f_0, f_{i\alpha}, \alpha \in \{\alpha\}_i, i = 1, \cdots, n) = (f_0, f)$ be a continuous $(N + 1)$-vector function on M, and assume that the set $Q(t, x) = \{z = (z^0, \cdots, z^n) : z^0 \geqq f_0(t, x, u), (z^1, \cdots, z^n) = f(t, x, u)$ for some $u \in U(t) = U(r)\}$ is a closed convex subset of E_n for every $(t, x) \in A$ and satisfies property (Q) *in A. Assume that $f_{i\alpha}(t, x, u) > -c_{i\alpha}$, $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n$, and $f_0(t, x, u) \geqq -M_0$ for all $(t, x, u) \in M$ and some constants $c_{i\alpha} \geqq 0$, $M_0 \geqq 0$.*

Assume further that: (Γ) *whenever (t, x, u_1), (t, x, u_2) are in M, and $f_{i\gamma}(t, x, u_1) = f_{i\gamma}(t, x, u_2)$ for all $\gamma \in \{\gamma\}_i$, $i = 1, \cdots, n$, then $f_{i\alpha}(t, x, u_1) = f_{i\alpha}(t, x, u_2)$ for all $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n$, and $f_0(t, x, u_1) = f_0(t, x, u_2)$* (see also Remark 3.2). *Also assume that*: (Δ) *for $\gamma \in \{\gamma\}_i$, $i = 1, \cdots, n$,*

$$f_{i\gamma}(t, x, u) = g_{i\gamma}(t, x) + w_{i\gamma}(s, x) k_{i\gamma}(r, u), \tag{5.1}$$

where $w_{i\gamma}(s, x) \neq 0$ for all (s, x) for which $(r, s, x) \in A$ for some $r \in R$.

Let $g(t, x) = (g_{i\gamma})$, $w(s, x) = (w_{i\gamma})$, $k(r, u) = (k_{i\gamma})$, $\hat{f}(t, y, u) = (f_{i\gamma})$, $\gamma \in \{\gamma\}_i$, $i = 1, \cdots, n$. For any fixed $s_1, s_2 \in S$, and $(r, s_1, x_1), (r, s_2, x_2) \in A$, let $Q(r, x_1, x_2; s_1, s_2) = \{z = (z^1, \cdots, z^{2K}) : (z^1, \cdots, z^K) = \hat{f}(r, s_1, x_1, u), (z^{K+1}, \cdots, z^{2K}) = \hat{f}(r, s_2, x_2, u)$ for some $u \in U(r)\}$. We assume that: (Ξ) *for any $s_1, s_2 \in S$, $Q(r, x_1, x_2; s_1, s_2)$ satisfies property* (Q) *for all (r, x_1, x_2) with (r, s_1, x_1), $(r, s_2, x_2) \in A$* (see also Corollaries 5.1 and 5.2).

Let (B) be a set of boundary conditions satisfying properties (P_1) *and* (P_2). *Let Ω be a nonempty complete class of admissible pairs $x(t), u(r)$, that is, pairs*

$$x(t) = (x^1(t), \cdots, x^n(t)), \qquad u(r) = (u^1(r), \cdots, u^m(r)),$$

$$t = (r, s) \in G, \quad x^i(t) \in W^{l_i}_{p_i}(G), \quad 1 \leqq l_i \leqq l, \quad p_i > 1, \quad i = 1, \cdots, n,$$

$u(r)$ measurable in R (a function of r alone, as stated in § 2), *satisfying*: (a) *the con-*

straints $(t, x(t)) \in A$, $u(r) \in U(r)$, *a.e. in* G *and* R, *respectively*, (b) *the system of partial differential equations* $D^{\alpha}x^{i}(t) = f_{i\alpha}(t, x(t), u(r))$ *a.e. in* G, $\alpha \in \{\alpha\}_i$, $i = 1, \cdots, n$, (c) *the boundary conditions* (B) *on the boundary* ∂G *of* G *involving the boundary values of* $x^{i}(t)$ *and* $D^{\beta}x^{i}(t)$, $0 \leqq |\beta| \leqq l_i - 1, i = 1, \cdots, n$.

Moreover, we ask that $x(t)$ *satisfy*: (d) *the system of inequalities*

$$\int_G |D^{\beta}x^{i}(t)|^{p_i}\, dt \leqq N_{i\beta}$$

for all β *with* $|\beta| = l_i, \beta \notin \{\alpha\}_i, i = 1, \cdots, n$, *where* $N_{i\beta}$ *are given constants*, (e) $f_0(t, x(t), u(r))$ *is of class* $L_1(G)$. *Assume further that*: (f) *if*

$$\int_G f_0(t, x(t), u(r))\, dt \leqq L_0$$

for some constant L_0 *and pairs* $x(t), u(r)$ *in* Ω, *then for the same pairs we also have*

$$\int_G |D^{\alpha}x^{i}(t)|^{p_i}\, dt \leqq L_{i\alpha}, \qquad \alpha \in \{\alpha\}_i, \quad i = 1, \cdots, n,$$

for some constant $L_{i\alpha}$ *depending only on* L_0, Ω, p_i, l_i, *and the boundary conditions* (B), *but not on the pair* x, u *itself.*

Then the cost functional

$$I[x, u] = \int_G f_0(t, x(t), u(r))\, dt$$

possesses an absolute minimum in Ω.

Proof. The proof of the existence theorem is essentially the same as that of Existence Theorem 2 of Cesari [6, § 5], with the only difference occurring at the end of the proof, where instead of applying Cesari's closure theorems, we apply the closure theorem of the present paper.

Remark 5.1. The domain G may be modified as described in Remark 3.3.

COROLLARY 5.1 (for compact control spaces) and COROLLARY 5.2. *Let the assumptions of the existence theorem be modified in precisely the manner that the assumptions of the closure theorem were modified in Corollary* 3.1 *or Corollary* 3.2, *respectively. Then the conclusions of the existence theorem remain valid.*

Proof. Since the closure theorem remains in force, so does the existence theorem.

Remark 5.2. The existence theorem also holds, with certain modifications, for G unbounded. Also, the assumption that $p_i > 1$ may be relaxed to $p_i \geqq 1$, $i = 1, \cdots, n$. For details, see the discussion of Cesari in [7] and [8] for usual multidimensional systems; the modifications are the same for our systems with lower-dimensional controls.

Remark 5.3. Let us modify the control systems so far studied by supposing that the control u may be written as $u = (v_1, v_2), v_1 = (u^1, \cdots, u^a), v_2 = (u^{a+1}, \cdots, u^m), u(t) = (v_1(r), v_2(t))$, that is, the first a components of u must depend upon r alone, while the remaining components of u may be functions of $t = (r, s)$. Also suppose $v_1(r) \in U_1(r), v_2(t) \in U_2(t)$, that is, the control set $U(t) = U_1(r) \times U_2(t)$. Then the above existence theorem, and all its corollaries, continue to hold if: (i) for $\gamma \in \{\gamma\}_i, i = 1, \cdots, n, f_{i\gamma}$ does not contain both v_1 and v_2; (ii) let $\{\theta\}_i$, $i = 1, \cdots, n$, be a new collection of indices consisting of those indices from $\{\gamma\}_i$

for which $f_{i\gamma}$ does not contain v_2; we ask that all the hypotheses of the existence theorem, and its corollaries, with the exception of assumption (Γ), hold with $\{\gamma\}_i$ replaced by $\{\theta\}_i, i = 1, \cdots, n$; (iii) we modify assumption (Γ) by asking that whenever $(t, x, v_1, v_2), (t, x, v_1', v_2)$ are in M, and $f_{i\theta}(t, x, v_1') = f_{i\theta}(t, x, v_1')$ for all $\theta \in \{\theta\}_i, i = 1, \cdots, n$, then $f_{i\alpha}(t, x, v_1, v_2) = f_{i\alpha}(t, x, v_1', v_2)$ for all $\alpha \in \{\alpha\}_i, i = 1, \cdots, n$, and $f_0(t, x, v_1, v_2) = f_0(t, x, v_1', v_2)$.

The proof of this statement is essentially the same as that of the existence theorem.

6. Examples

Example 1 (Cesari [5, § 5]). We wish to determine $u = u(\xi), v = (\xi, \eta)$, so as to minimize

$$I[x, u, v] = \int_G [A(\xi, \eta, x)u^2 + 2B(\xi, \eta, x)uv + C(\xi, \eta, x)v^2]\, d\xi\, d\eta,$$

subject to

$$\frac{\partial x}{\partial \xi} = a_1(\eta)a_2(\xi)u, \quad \frac{\partial x}{\partial \eta} = c(\xi, \eta)u + d(\xi, \eta)v,$$

where the domain G is an open set in the $\xi\eta$-space of class K_1, with meas $\{\eta : G_\eta = R\} > 0, R = \{\xi : (\xi, \eta) \in G$ for some $\eta \in E_1\}$ (see also Remark 3.3). We assume that A, B, C are continuous functions of $(\xi, \eta, x) \in \text{cl}\, G \times E_1, A > 0, C > 0$, $B^2 < AC$, that a_1, a_2, c, d are continuous functions of (ξ, η) on cl G, and that $a_1(\eta), a_2(\xi) \neq 0$ for $(\xi, \eta) \in G$. Let $(u, v) \in U = \{u : |u| \leqq L, |v| \leqq L\}$, and assume that boundary conditions are given by $x = \Phi$ on ∂G, where Φ is a given continuous function on ∂G. Let Ω consist of the pairs $x(\xi, \eta), (u(\xi), v(\xi, \eta)), (\xi, \eta) \in G$, satisfying $x \in W_2^1(G)$, u measurable in R, v measurable in G, and the above conditions, and suppose $\Omega \neq \varnothing$.

By the above assumptions, there is some M so that $|x(\xi, \eta)| \leqq M$; let $A = \text{cl}\, G \times [-M, M]$. Let M_0 be the maximum value of $Au^2 + 2Buv + Cv^2$ for $(\xi, \eta, x) \in A, (u, v) \in U$. Then the set $\tilde{\tilde{Q}}(\xi, \eta, x; M_0) = \{(z^0, z^1, z^2) : Au^2 + 2Buv + Cv^2 \leqq z^0 \leqq M_0, z^1 = a_1a_2u, z^2 = cu + dv\}$ is convex, and hence upper semicontinuous [5, § 4]; also $k(\xi, U)$ satisfies Corollary 3.1. By Corollary 5.1 of the existence theorem, and Remarks 3.1 and 5.3, it follows that an optimal pair in Ω exists for this system.

Example 2 (modification of the Neyman–Pearson lemma [11]). We wish to determine $u(x)$ so as to maximize

$$I = \int_{a_1}^{a_2} \int_{b_1}^{b_2} u(x)g(x, y)\, dx\, dy,$$

subject to

$$\int_{a_1}^{a_2} \int_{b_1}^{b_2} u(x)f(x, y)\, dx\, dx \leqq C,$$

with $0 \leqq u \leqq 1$. Here $G = [a_1, a_2] \times [b_1, b_2]$, and f and g are nonnegative, bounded, continuous functions on G, with $\int_G f, \int_G g$ both one, and $C \geqq 0$, a constant.

Let us first assume a_1, a_2, b_1 and b_2 are finite. In order to apply the existence

theorem, let us rewrite this system in Rashevsky form. To this end, let

$$z^1(x, y) = \int_{a_1}^{x} \int_{b_1}^{y} u(x) f(x, y)\, dx\, dy,$$

and let

$$z_x^1 = z^2, \quad z_y^1 = z^3.$$

Then the above system may be written:

Minimize

$$\int_{a_1}^{a_2} \int_{b_1}^{b_2} u(x)[-g(x, y)]\, dx\, dy,$$

subject to

$$\begin{aligned} z_x^1 &= z^2, \quad z_x^3 = u(x) f(x, y), \\ z_y^1 &= z^3, \quad z_y^2 = u(x) f(x, y), \end{aligned} \tag{6.1}$$

with $z = (z^1, z^2, z^3) \in W_2^1(G)$, with control $u(x)$ a measurable function in $[a_1, a_2]$, $0 \leqq u \leqq 1$, with ξ_1, ξ_2 arbitrary, and boundary conditions $0 = z^1(a_1, y) = z^1(x, b_1) = z^2(x, b_1) = z^3(a_1, y), (x, y) \in G$, and $z^1(a_2, b_2) \leqq C$.

We require that $\iint_G |z_x^2|^2\, dx\, dy \leqq M_1, \iint_G |z_y^3|^2\, dx\, dy \leqq M_2$ for some constants M_1, M_2. Referring to relation (5.1) and the assumption that $w(s, x) \neq 0$, we also assume $f(x, y) > 0$. Assumptions (Δ) and (Γ) are thus satisfied. Moreover, since $z^1(x, y)$ is continuous in G, the given boundary conditions also satisfy conditions (P_1) and (P_2). In addition, the control set is $U = [0, 1]$, a compact fixed subset of E_1, and $Q(x, y, z)$ is a convex set since the controls enter linearly in (6.1). Also, $f_0(x, y, z, u) \geqq -g(x, y) \in L_1(G)$, and

$$\int_{a_1}^{a_2} \int_{b_1}^{b_2} |u(x) g(x, y)|\, dx\, dy \leqq 1.$$

Again, from the bounds on f, g and u, there is some constant M such that $|z^i| \leqq M$, $i = 1, 2, 3$, so that we can set $A = \operatorname{cl} G \times [-M, M]^3$, a compact set in E_4. Hence, by Corollary 5.1, if the system is controllable, then there exists an optimal pair $z(x, y), u(x)$ for this system. Since $u(x) \equiv 0$ is an admissible control, an optimal pair thus exists. It can also be shown that under minor changes of the above assumptions, we can also conclude that an optimal pair exists if we allow any of a_1, a_2, b_1, b_2 to be infinite (see Remark 5.2 and [8]).

Remark 6.1. For the usual form of the Neyman–Pearson lemma, in which $u = u(x, y)$, a function of both x and y, the results of Cesari in [7] may be directly applied to show the existence of optimal pairs.

Example 3 (search for a diffusing target [9]). We wish to determine $u(y)$ so as to maximize

$$\int_{a_1}^{a_2} \int_{b_1}^{b_2} g(x, u(y)) p(x, y)\, dx\, dy,$$

subject to $0 \leqq u$, and

$$\frac{\partial p}{\partial y} = a \frac{\partial^2 p}{\partial x^2} - b \frac{\partial p}{\partial x} + p\left[\int_{a_1}^{a_2} g(x, u(y)) p(x, y)\, dx - g(x, u(y))\right]. \tag{6.2}$$

Here, $G = [a_1, a_2] \times [b_1, b_2]$, and g is a nonnegative, bounded, continuous function of x and u. Let us first assume G is bounded. Again, writing this system in Rashevsky form, we obtain:

Minimize

$$\int_{a_1}^{a_2} \int_{b_1}^{b_2} [-g(x, u(y))] z^4(x, y)\, dx\, dy,$$

subject to

$$z_x^1 = z^2, \quad z_x^3 = gz^4, \quad z_x^4 = z^5,$$

$$z_y^1 = z^3, \quad z_y^2 = gz^4, \quad z_y^4 = az^6 - bz^5 + z^4 z^3(a_2, y) - z^4 g, \quad z_x^5 = z^6, \tag{6.3}$$

with $z = (z^i)$, $i = 1$ to 6, in $W_2^1(G)$, with control $u(y)$, measurable in $[b_1, b_2]$, $u \geqq 0$, with ξ_i, $i = 1$ to 5 arbitrary, and with boundary conditions $0 = z^1(a_1, y) = z^1(x, b_1) = z^2(x, b_1) = z^3(a_1, y)$, and suitable boundary conditions on $z^4 = p$; typically, we specify values for $p(0, x)$ and /or $p(y, 0)$. Again, $\{\alpha\}_i = \{(1, 0), (0, 1)\}$, $\{\gamma\}_i = \{(0, 1)\}$, $i = 1$ to 6. We assume that

$$g(x, u) = g_1(x) + g_2(x) k(u),$$

with $z^4 = p > 0$, $g_2(x) \neq 0$, and $k(u)$ a bounded, upper semicontinuous function of u, and $k(U)$ a convex set in E_1, $U = [0, \infty)$. Here $p(x, y)$ represents a probability density function which is associated with a moving target, and which depends upon u according to the diffusion equation (6.2) (see [9]). Thus it follows that the assumption $p > 0$ will be satisfied in many cases.

Alternately, we may consider only those pairs of functions z, u, so that $z^4 = p \geqq \delta > 0$, for some $\delta > 0$. Assumptions (Δ) and (Γ) are then satisfied. Moreover, the given boundary conditions also satisfy conditions (P_1) and (P_2). In addition, the control set $U = [0, \infty)$ is a closed fixed subset of E_1, and $A = G \times E_6$, a closed subset in E_8. We take as Ω, the class of admissible pairs, those functions $z(x, y), u(y)$, $z = (z^i)$, $i = 1$ to 6, which satisfy the above conditions, and for which

$$\iint_G |D_x z^i|^2\, dx\, dy \leqq N_{i1}, \quad i = 2, 6; \quad \int_G |D_y z^i|^2\, dx\, dy \leqq N_{i2}, \quad i = 3, 5, 6. \tag{6.4}$$

Since $z_{xy}^4 = D_y z^5$, and by (6.4), there is an N so that $|z^4| \leqq N$. Hence $f_0 = -g(x, u) \cdot p(x, y) \geqq -M$ for some constant M. Finally, suppose $u(y) \equiv 0$ is an admissible control for this system. Hence, by Corollary 5.2, an optimal pair exists for this system. Again, since $p(x, y)$ represents a probability density, we expect $\iint_G p(x, y)\, dx\, dy = 1$ for any $u = u(y)$; in particular, $p \in L_1(G)$. In this case $f_0 = -g(x, u) p(x, y) \geqq -Bp(x, y)$, where $|g(x, u)| \leqq B$; hence, it may again be shown that an optimal pair exists if we allow any of the a_1, a_2, b_1, b_2 to be infinite.

Remark 6.2. Since z^6 enters (6.3) linearly, it may be shown that we need only assume that $z^i \in W_2^1(G)$, $i = 1$ to 5, and $z^6 \in L_2(G)$. Further reductions in the assumptions may also be made. These points will be discussed elsewhere.

Example 4 (stochastic multidimensional systems). Let us consider a usual multidimensional control system, in which the control u is a function of t, and $x \in W_{p_i}^{l_i}(G)$. Suppose, however, that due to the presence of stochastic disturbances,

or "noise", the state equation now contains a perturbation, or stochastic process $\eta(t, a), t \in G, a \in I \subset E_d, \eta \in E_n$. Assume further that for some reason, such as lack of information, the control u must remain a function of t alone (open loop). Hence, for a.a. $(t, a) \in \tilde{G} = G \times I$, we have

$$D^{\alpha}x^i(t, a) = f_{i\alpha}(t, x(t, a), u(t), \eta(t, a)),$$

where $\alpha \in \{\alpha\}_i$ denotes differentiation with respect to $(t^1, \cdots, t^v)$ only. Suppose moreover that $\eta(t, a)$ is sufficiently well-behaved so that $(x^i, \eta^i) \in W^{l_i}_{p_i}(\tilde{G})$. Such situations can arise, for instance, if η satisfies an equation similar to a diffusion equation, or if η satisfies a system of partial differential equations containing a random variable (see [1] for the analogous situation for one-dimensional systems).

We now have a multidimensional system as described in § 2, with G replaced by $\tilde{G}$, t replaced by (t, a), r replaced by t, s replaced by a, R replaced by G, x replaced by (x, η), n replaced by $2n$, and $\{\gamma\}_i = \{\alpha\}_i, i = 1, \cdots, N$. For each admissible pair (x, u), we associate the cost functional

$$I[x, u] \qquad h(t, x(t, a), u(t))\, dt\, da,$$

where h now incorporates the cost associated with the stochastic deviations due to η. For instance, let $p(a)$ be a probability density associated with $a \in I$. Then the cost may be taken as the expected value of $\int_G f_0$, or

$$I[x, u] = \int_I \int_G f_0(t, x(t, a), u(t))p(a)\, dt\, da.$$

The results of this paper can now be applied to these stochastic multidimensional control systems.

Acknowledgments. The author wishes to thank Lamberto Cesari for his many helpful comments and assistance in the preparation of this paper, and Stephen Pollock for his discussions on the search problem.

REFERENCES

[1] R. F. Baum, *An existence theorem for optimal control systems with state variable in C, and stochastic control problems*, J. Optimization Theory Appl., 5 (1970), pp. 335–345.

[2] ———, *Necessary conditions for multidimensional control systems with lower dimensional control*, Tech. Rep. 71-2, Department of Industrial Engineering, University of Michigan, Ann Arbor, 1971.

[3] A. G. Butkovsky, *Distributed Control Systems*, American Elsevier, New York, 1969.

[4] 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.

[5] ———, *Existence theorems for multidimensional problems of optimal control*, Differential Equations and Dynamical Systems, J. K. Hale and J. P. LaSalle, eds., Academic Press, New York, 1967, pp. 115–132.

[6] ———, *Existence theorems for multidimensional Lagrange problems*, J. Optimization Theory Appl., 1 (1967), pp. 87–112.

[7] ———, *Sobolev spaces and multidimensional Lagrange problems of optimization*, Ann. Scuola Norm. Sup., 15A aa (1968), pp. 194–227.

[8] ———, *Multidimensional Lagrange problems of optimization in a fixed domain and an application to a problem of magnetohydrodynamics*, Arch. Rational Mech. Anal., 29 (1968), pp. 81–104.

[9] O. HELLMAN, *On the optimal search for a randomly moving target*, SIAM J. Appl. Math., to appear.
[10] S. L. SOBOLEV, *Applications of functional analysis in mathematical physics*, Translations Math. Monographs, vol. 7, American Mathematical Society, Providence, R.I., 1963.
[11] D. H. WAGNER, *Nonlinear functional versions of the Neyman–Pearson lemma*, SIAM Rev., 11 (1969), pp. 52–65.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

ON THE STABILITY OF DIFFERENTIAL EQUATIONS*

JOHN J. H. MILLER†

Abstract. A polynomial is said to be of type (p_1, p_2, p_3) relative to any directed line in the complex plane if it has p_1 zeros to the left of, p_2 on, and p_3 to the right of the line. Stability criteria for differential equations often involve checking that a polynomial or a family of polynomials is of type $(p_1, p_2, 0)$ or of type $(p_1, 0, 0)$ relative to the imaginary axis. Here we reconsider the practical problem of showing that a polynomial is of one or other of these types relative to any line through the origin, and we establish that the testing of a polynomial of degree n may always be reduced to the testing of one of degree $n - 1$. Repeated application of this result reduces the problem to a trivial one. There is a long history of similar results in the case $p_2 = 0$, particularly for polynomials with real coefficients. Our main contribution is to extend the techniques when $p_2 \neq 0$.

Our results are valid for polynomials with real or complex coefficients, they are of an algorithmic nature and they reveal the nature and importance of self-inversiveness.

1. Introduction. Frequently the question of the stability of a system of differential equations may be reduced to the problem of determining whether or not the zeros of a polynomial all lie in the open left-half-plane. Necessary and sufficient conditions for a polynomial to have this property have been found by many authors. One obvious approach is to find *numerical approximations* to all roots. However, we are interested here in the *exact algebraic* approach initiated independently by Routh [7] and Hurwitz [6]. This approach is essential if the coefficients of the polynomial depend on parameters and we have to determine the values these may take without destroying the stability of the polynomial.

In practice it is recommended that all criteria of algebraic type be checked either by exact hand computation or by an *algebraic* implementation on a computer. It is well known that the locations of the zeros of a polynomial are often extremely sensitive to small perturbations of the coefficients (see, for example, Wilkinson [8]). For this reason we question the validity of the *numerical* computer implementation of Routh–Hurwitz-type tests, which we have seen from time to time.

Our starting point is a paper by Duffin [3], where an algorithm is obtained which determines whether or not a polynomial of degree n with real or complex coefficients is stable. This is done by reducing the problem to the same question for a polynomial of degree $n - 1$. After a sufficient number of reductions the problem is clearly trivial. The only restriction in Duffin's algorithm is that the constant term of each polynomial must be real, and his algorithm guarantees that if the original polynomial has this property, then the reduced polynomials have also.

In this paper we are concerned with more general problems of the same type. Our main contribution is to show what modifications have to be made to the standard techniques when the polynomials are allowed to have zeros on the imaginary axis, in addition to having zeros in the open half-planes on either side of it. For convenience only we develop the theory for an arbitrary line through

* Received by the editors June 29, 1971, and in final revised form April 6, 1972.
† School of Mathematics, Trinity College, University of Dublin, Dublin 2, Ireland.

the origin rather than just the imaginary axis. This provides us with analogous results for the real axis, for example, without any additional work. A trivial extension using translation and rotation would give similar results for an arbitrary line in the complex plane. In § 2 we introduce various classes of stable polynomials and we indicate briefly some areas where they arise. In § 3 we discuss inversions of points and polynomials in a line, and we introduce the idea of a self-inversive polynomial. Reduced polynomials are defined in § 4, and some of their properties are established. Our main theoretical results are proved in § 5, and in § 6 these are used to prove our reduction algorithms. Finally in § 7 we present some examples.

It should be noted that while our algorithms apply to more general problems than does Duffin's, they do *not* reduce to his for the simpler problem, and for this problem his are easier to apply. However, in a subsequent paper we hope to establish a generalization of Duffin's algorithm which will apply to our more general case.

2. Classes of polynomials. Given any directed line L in the complex plane we say that a polynomial of degree n is of *type* (p_1, p_2, p_3) *relative to* L, where $p_1 + p_2 + p_3 = n$, if it has p_2 zeros on L and p_1 (p_3) zeros on the left (right) of L. If L is the imaginary axis, we have the special case treated by Routh and Hurwitz, and if L is the real axis, that of Hermite.

We introduce now some important classes of polynomials. The first is the largest and contains all the subsequent ones. This is the class of *Petrowsky–Gårding polynomials*: those polynomials of degree n which are of type $(p, n - p, 0)$ for some nonnegative integer p. These occur in the study of weakly well-posed Cauchy problems for partial differential operators. The subclass of these of type $(n, 0, 0)$ are the *Hurwitz polynomials*, which are important in the theory of stable dynamical systems where the effect of a small perturbation becomes negligible after a finite period of time. Another important subclass are the polynomials of type $(0, n, 0)$, which are called the *conservative polynomials*. These arise in conservative physical systems, and in the special case where L is the real axis they coincide with the hyperbolic polynomials introduced by Gårding [5]. It is often important to know whether or not a zero on L is repeated. This leads us to define the *simple Petrowsky–Gårding polynomials* as those polynomials which are Petrowsky–Gårding and have only simple zeros on L. The class of *simple conservative polynomials* is defined analogously.

3. Inversion and self-inversive polynomials. For simplicity we restrict our discussion to a line L_θ through the origin, making an angle θ with the real axis, $0 \leqq \theta < 2\pi$, and directed in the sense that L_0 is the real axis. It is an easy matter to generalize all the subsequent definitions and results to an arbitrary line L.

For any point z in the complex plane we define its *inverse* z^* in L_θ by

$$z^* = e^{2i\theta}\bar{z}. \tag{3.1}$$

This is simply the reflection of z in L_θ, so that $z^{**} = z$ and $z \in L_\theta$ if and only if $z = z^*$. Moreover, if f is a polynomial, we define its inverse f^* in L_θ to be the polynomial

$$f^*(z) = \overline{f(z^*)}. \tag{3.2}$$

Thus if $f(z) = \sum_{j=0}^{n} a_j z^j$, we have

$$f^*(z) = \sum_{j=0}^{n} \overline{a_j} e^{-2ij\theta} z^j$$

which is of the same degree as f. Furthermore,

$$f^{**} = f, \quad (fg)^* = f^* g^*, \quad (f+g)^* = f^* + g^*, \quad \left(\frac{f}{g}\right)^* = \frac{f^*}{g^*},$$

and $(\lambda f)^* = \bar{\lambda} f^*$ for any polynomials f, g and any scalar λ.

Consider now any $\alpha \in L_\theta$; then $\alpha^* = \alpha$ and from (3.2), $f^*(\alpha) = \overline{f(\alpha)}$ so that

$$|f^*(\alpha)| = |f(\alpha)| \quad \text{for all } \alpha \in L_\theta. \tag{3.3}$$

Also from (3.2) we see that z_j is a zero of f if and only if z_j^* is a zero of f^* with the same multiplicity. Thus f is of type (p_1, p_2, p_3) if and only if f^* is of type (p_3, p_2, p_1), and the p_2 zeros on L_θ are common to both f and f^*.

We now introduce an analogue of a concept due to Bonsall and Marden [2]. We say that f is *self-inversive* (with respect to L_θ) if f and f^* have the same set of zeros and the multiplicity of each distinct zero is the same in both f and f^*. Alternative criteria for self-inversiveness are contained in the following theorem.

THEOREM 3.1. *The following conditions are equivalent*:

(a) *f is self-inversive.*

(b) *The zeros of f and their multiplicities are symmetric in L_θ.*

(c) *$f^*(w)f(z) = f(w)f^*(z)$ for all $z, w \in \mathbb{C}$.*

(d) *$|f^*(z)| = |f(z)|$ for all $z \in \mathbb{C}$.*

Proof. The equivalence of (a) and (b) is easy. Suppose now that (a) holds. Then for some constant c, $f^*(z) \equiv cf(z)$ so that if w is not a zero of f, then $c = f^*(w)/f(w)$. On the other hand, if w is a zero of f, it is also a zero of f^*. In either case (c) holds. The converse, (c) implies (a), is trivial. Choosing $w = \alpha \in L_\theta$ in (c) and using (3.3) gives (d). To conclude the proof it suffices to show that (d) implies (a). Certainly (d) implies that f and f^* have the same set of zeros, and it remains only to show that the multiplicities are the same in both f and f^*. Suppose z_j is a zero of f of multiplicity m_j. Then from (d), z_j must also be a zero of f^* of multiplicity n_j say. Thus $f(z) = (z - z_j)^{m_j} g(z)$, $f^*(z) = (z - z_j)^{n_j} h(z)$ for some g and h such that $g(z_j) \neq 0, h(z_j) \neq 0$. Assume now that $m_j > n_j$; then (d) implies that

$$|z - z_j|^{m_j - n_j} |g(z)| = |h(z)| \quad \text{for } z \in \mathbb{C}.$$

Letting $z = z_j$ we see that $h(z_j) = 0$, a contradiction. Similarly we arrive at a contradiction if $m_j < n_j$. Hence $m_j = n_j$ and since this holds for each j we have established (a).

COROLLARY 3.2. *Suppose f is a self-inversive polynomial of degree n; then f is of type $(p, n - 2p, p)$ for some nonnegative integer p.*

Proof. The proof is immediate from (b).

Taking $\theta = 0$ we note that a polynomial is self-inversive with respect to L_0 if and only if it has real coefficients. Hence the class of self-inversive polynomials is the analogue for $\theta \neq 0$ of the class of polynomials with real coefficients for $\theta = 0$.

4. Reduction of polynomials. For any polynomial f we define the *reduced polynomial* $\check{f}$ by

$$\check{f}(z) = \frac{f^*(w)f(z) - f(w)f^*(z)}{z - w}, \tag{4.1}$$

where $w \in \mathbb{C}$ is regarded as a parameter. We assume throughout that $w \notin L_\theta$, and that it is not a zero of f or of f^*. Since w is a zero of the numerator, $\check{f}$ is a polynomial in z of degree at most $n - 1$. On the other hand, the coefficient of z^{n-1} is $a_n f^*(w) - e^{-2in\theta}\bar{a}_n f(w)$, so that f is of degree $n - 1$ if $|f^*(w)| \neq |f(w)|$. By Theorem 3.1 (c), f is self-inversive if and only if $\check{f} \equiv 0$. It is easy to see that if $f = gh$, then

$$\check{f}(z) = g^*(w)g(z)\check{h}(z) + h^*(z)h(w)\check{g}(z). \tag{4.2}$$

In particular, if $f = g\psi$, where ψ is self-inversive, then by (4.2) and Theorem 3.1 (c),

$$\check{f}(z) = \psi^*(z)\psi(w)\check{g}(z) = \psi^*(w)\psi(z)\check{g}(z). \tag{4.3}$$

Thus if f is not self-inversive, every self-inversive factor of f is a factor of $\check{f}$. Furthermore, by (3.4) we have

$$|f^*(w)| - |f(w)| = |\psi^*(w)g^*(w)| - |\psi(w)g(w)| = |\psi(w)|(|g^*(w)| - |g(w)|), \tag{4.4}$$

so that $|f^*(w)| - |f(w)|$ and $|g^*(w)| - |g(w)|$ are either both zero or have the same sign.

Suppose now that $\check{f} = \varphi h$, where φ is self-inversive. Then $(\check{f})^* = \varphi^* h^*$. In view of (3.1) and (4.1) these equations may be written in the form

$$f^*(w)f(z) - f(w)f^*(z) = (z - w)\varphi(z)h(z),$$
$$\overline{f^*(w)}f^*(z) - \overline{f(w)}f(z) = e^{-2i\theta}(z - w^*)\varphi^*(z)h^*(z).$$

Multiplying the first by $\overline{f^*(w)}$, the second by $f(w)$ and adding, we obtain, by (3.2) and Theorem 3.1 (c),

$$(|f^*(w)|^2 - |f(w)|^2)f(z) = \frac{\varphi(z)}{\varphi(w)}[(z - w)\varphi(w)f(w^*)h(z) + e^{-2i\theta}(z - w^*)\varphi^*(w)f(w)h^*(z)].$$

From this we conclude that if $|f^*(w)| \neq |f(w)|$, then φ is a factor of f. We have established our next theorem.

THEOREM 4.1. *Assume f is not self-inversive and suppose that $f = \psi g$, where ψ is the maximal self-inversive factor of f. Then ψ is a factor of $\check{f}$, and $|f^*(w)| - |f(w)|$, $|g^*(w)| - |g(w)|$ are either both zero or have the same sign. If $|f^*(w)| \neq |f(w)|$, then ψ is the maximal self-inversive factor of $\check{f}$.*

From this and Corollary 3.2 our next result follows immediately.

COROLLARY 4.2. *Assume $|f^*(w)| \neq |f(w)|$ and that the maximal self-inversive factor ψ of f is of degree m. Then f is of type (p_1, p_2, p_3) if and only if g is of type $(p_1 - q, 0, p_3 - q)$, where $q = \frac{1}{2}(m - p_2)$.*

We remark that if f is not self-inversive, then by Theorem 3.1 (d) it is always possible to find a w such that

$$|f^*(w)| \neq |f(w)|.$$

Since $\check{f}$ vanishes identically if f is self-inversive, we now take up the problem of obtaining a nontrivial reduction of a self-inversive polynomial. Our method is to replace f by a polynomial $\tilde{f}$, whose reduction is nontrivial. Naturally we need to know the relation between the type of f and that of $\tilde{f}$.

For any polynomial of degree n we define

$$\tilde{f}(z) = f(z) + e^{-2i\theta}\xi(z - w^*)(nf(z) - (z - w)f'(z)), \tag{4.5}$$

where ξ is for the moment an arbitrary complex parameter. Noting that $(f')^* = e^{2i\theta}(f^*)'$, it is easy to see that

$$(\tilde{f})^*(z) = f^*(z) + e^{-2i\theta}\xi^*(z - w)(nf^*(z) - (z - w^*)(f^*)'(z)) \tag{4.6}$$

and thus

$$\tilde{f}(w) = (1 + e^{-2i\theta}n\xi(w - w^*))f(w), \qquad (\tilde{f})^*(w) = f^*(w). \tag{4.7}$$

It is also clear that, for all sufficiently small $|\xi|$, $\tilde{f}$ and $(\tilde{f})^*$ are of degree n.

LEMMA 4.3. *Assume f is self-inversive, ξ is to the left of L_θ, and w is to the left (right) of L_θ. Then for all sufficiently small $|\xi|$,* (a) $|(\tilde{f})^*(w)| - |\tilde{f}(w)| > 0$ (<0), (b) $(\tilde{f})^\vee$ *is of degree $n - 1$ and* (c) $(\tilde{f})^\vee(z)$ *differs from $nf(z) - (z - w^*)f'(z)$ by only a nonzero constant factor.*

Proof. Since f is self-inversive, $|f^*(w)| = |f(w)|$. Hence from (4.7),

$$|(\tilde{f})^*(w)|^2 - |\tilde{f}(w)|^2 = (1 - |1 + e^{-2i\theta}n\xi(w - w^*)|^2)|f(w)|^2.$$

But

$$|1 + e^{-2i\theta}n\xi(w - w^*)|^2 = 1 + 2n\,\mathrm{Re}\,(e^{-2i\theta}\xi(w - w^*)) + n^2|\xi|^2|w - w^*|^2,$$

and for ξ to the left of L_θ, it is easy to see that $\mathrm{Re}\,(e^{-2i\theta}\xi(w - w^*)) < 0$ (>0) for w to the left (right) of L_θ. We conclude therefore that $|1 + e^{-2i\theta}n\xi(w - w^*)|^2 < 1$ (>1) for all sufficiently small $|\xi|$ when w is to the left (right) of L_θ.

Thus (a) holds and, as we pointed out above, this implies (b).

To establish (c) we note that $e^{-2i\theta}\xi\xi^* = |\xi|^2$, and since f is self-inversive, $f^*(w)f(z) = f(w)f^*(z)$, $f^*(w)f'(z) = f(w)(f^*)'(z)$. Thus from (4.5), (4.6) and (4.7),

$$\begin{aligned}(\tilde{f})^\vee(z) &= \frac{(\tilde{f})^*(w)\tilde{f}(z) - \tilde{f}(w)(\tilde{f})^*(z)}{z - w} \\ &= e^{-2i\theta}[(\xi - \xi^*) - n|\xi|^2(w - w^*)]f^*(w)[nf(z) - (z - w^*)f'(z)]\end{aligned}$$

from which (c) follows immediately.

We now give the relation between the type of f and that of $\tilde{f}$.

THEOREM 4.4. *Let f be any polynomial of dgree n with k distinct zeros on L_θ, $0 \leqq k \leqq n$, and let $\xi \in L_{\theta+\pi/2}$. Then for all sufficiently small $|\xi|$, f is of type (p_1, p_2, p_3) if and only if $\tilde{f}$ is of type $(p_1 + k, p_2 - k, p_3)$ for ξ to the left of L_θ and of type $(p_1, p_2 - k, p_3 + k)$ for ξ to the right of L_θ. Furthermore, the $p_2 - k$ zeros of $\tilde{f}$ on L_θ are a subset of the p_2 zeros of f on L_θ.*

Proof. Since the zeros of a polynomial are continuous functions of the coefficients and the coefficients of $\tilde{f}$ are small perturbations of those of f, it follows that every zero of f on one side of L_θ gives rise to a zero of $\tilde{f}$ on the same side of L_θ. It suffices therefore to determine what happens to a zero of f lying on L_θ.

Suppose then that

$$f(z) = (z - \alpha)^m h(z), \qquad \alpha \in L_\theta, \quad m \geqq 1, \quad h(\alpha) \neq 0.$$

Then

$$\tilde{f}(z) = (z - \alpha)^{m-1}[q(z)h(z) + \xi(z - \alpha)k(z)],$$

where

$$q(z) = z - \alpha - e^{-2i\theta} m\xi(z - w)(z - w^*),$$

$$k(z) = e^{-2i\theta}(z - w^*)(nh(z) - (z - w)h'(z)).$$

Thus α is a zero of $\tilde{f}$ of multiplicity $m - 1$, and the only other zero of $\tilde{f}$ corresponding to the zero α of f is the zero of $q(z)h(z) + \xi(z - \alpha)k(z)$ which converges to α as ξ goes to zero.

Since $h(\alpha) \neq 0$ there exists an open neighborhood N of α in which h does not vanish. The only zero of $q(z)h(z)$ in N is therefore the zero of $q(z)$ which converges to α as ξ goes to zero. This is easily seen to be

$$z_0(\xi) = \alpha + 2r\xi + O(\xi^2) \quad \text{as } \xi \to 0,$$

where $r = m(w - \alpha)^2/2$ is independent of ξ.

By hypothesis, ξ lies on $L_{\theta+\pi/2}$. Hence for all sufficiently small ξ, $z_0(\xi)$ lies on the same side of L_θ as ξ and its distance to both L_θ and the boundary of N is greater than $r|\xi|$.

We now consider the quotient of $\xi(z - \alpha)k(z)$ by $q(z)h(z)$ on a circle C about $z_0(\xi)$ of radius $r|\xi|$. The circle C certainly lies entirely within N and is on the same side of L_θ as ξ. Thus $k(h)/h(z)$ is uniformly bounded on C. Also it is easy to see that for all z on C,

$$|z - \alpha| \leqq |\xi|(3r + O(\xi)) \qquad \text{as } \xi \to 0,$$

$$|q(z)| \geqq r|\xi|(1 - O(\xi)) \qquad \text{as } \xi \to 0,$$

since $q(z_0) = 0$. We conclude therefore that for all sufficiently small ξ,

$$|\xi(z - \alpha)k(z)/q(z)h(z)| \leqq C|\xi| < 1$$

since C is a constant independent of ξ. Applying Rouché's theorem, we see that $q(z)h(z)$ and $q(z)h(z) + \xi(z - \alpha)k(z)$ have the same number of zeros within C, that is, precisely one. We have shown therefore that for each zero $\alpha \in L_\theta$ of f having multiplicity m, $\tilde{f}$ has a zero α of multiplicity $m - 1$ and a simple zero on the same side of L_θ as ξ. The theorem follows directly.

5. Main theorems. We begin by stating and proving a well-known variant of Rouché's theorem.

LEMMA 5.1. *Let g be any polynomial of type $(p_1, 0, p_3)$. Then the polynomial $g + \lambda g^*$ is of type $(p_1, 0, p_3)$ for all $|\lambda| < 1$ and of type $(p_3, 0, p_1)$ for all $|\lambda| > 1$.*

Proof. The zeros of $g + \lambda g^*$ are continuous functions of λ. Thus a zero leaving the left side of L_θ must cross L_θ at some point α. But then

$$g(\alpha) + \lambda g^*(\alpha) = 0,$$

and by (3.3) this implies that $|\lambda| = 1$. Thus the type remains constant for all $|\lambda| < 1$ and similarly for all $|\lambda| > 1$. Letting $\lambda = 0$ and $|\lambda| \to \infty$ respectively we obtain the desired result.

In connection with this we consider also the case where $|\lambda| = 1$. If $f = \psi g$, where ψ is the maximal self-inversive factor of f, then it is clear that

$$f(z) + \lambda f^*(z) = \frac{\psi(z)}{\psi(w)}[\psi(w)g(z) + \lambda\psi^*(w)g^*(z)].$$

Therefore ψ is a factor of $f + \lambda f^*$ provided the expression in brackets is not identically zero. Thus if ψ is self-inversive, but not conservative, it is clear that not all the zeros of $f + \lambda f^*$ lie on L_θ if $|\lambda| = 1$. We see, therefore, that the first statement of footnote 3 in [4, p. 272] is false. A particular counterexample for $\theta = \pi/2$ and $\lambda = 1$ is given by $f(z) = (z - 2)(z + 2)$, in which case $f(z) + \lambda f^*(z) = 2f(z)$.

THEOREM 5.2. *Suppose f is a polynomial of degree n and $|f^*(w)| \neq |f(w)|$. Then f is of type (p_1, p_2, p_3) if and only if $\check{f}$ is of type*

$$(p_1 - 1, p_2, p_3) \quad \text{for } |f^*(w)| > |f(w)| \quad \text{and } w \text{ left of } L_\theta,$$

$$(p_1, p_2, p_3 - 1) \quad \text{for } |f^*(w)| > |f(w)| \quad \text{and } w \text{ right of } L_\theta,$$

$$(p_3 - 1, p_2, p_1) \quad \text{for } |f^*(w)| < |f(w)| \quad \text{and } w \text{ left of } L_\theta,$$

$$(p_3, p_2, p_1 - 1) \quad \text{for } |f^*(w)| < |f(w)| \quad \text{and } w \text{ right of } L_\theta.$$

Furthermore, the zeros of f and $\check{f}$ on L_θ and their multiplicities are the same.

Proof. From Theorem 4.1 we see that the last statement of the theorem is true, and furthermore we need consider only polynomials g with no self-inversive factor and such that $|g^*(w)| \neq |g(w)|$. For such polynomials,

$$\check{g}(z) = \frac{g^*(w)g(z) - g(w)g^*(z)}{z - w}$$

is of degree one less than g and clearly $\check{g}$ has one zero less than $g^*(w)g - g(w)g^*$ on the same side of L_θ as w. The theorem follows directly by applying Lemma 5.1 to the polynomial $g + \lambda g^*$ with $\lambda = -g(w)/g^*(w)$.

From Theorem 3.1(d) we know that if f is not self-inversive, then we can find a w for which the hypothesis of Theorem 5.2 holds, but if f is self-inversive, then it is impossible. To deal with the self-inversive case we therefore need the following theorem.

THEOREM 5.3. *Suppose f is a self-inversive polynomial of degree n, with k distinct zeros on L_θ. Then f is of type $(p, n - 2p, p)$ if and only if $nf(z) - (z - w^*)f'(z)$ is of type*

$$(p + k - 1, n - 2p - k, p) \quad \text{for } w \text{ left of } L_\theta,$$

$$(p + k, n - 2p - k, p - 1) \quad \text{for } w \text{ right of } L_\theta.$$

Furthermore the $n - 2p - k$ zeros on L_θ are a subset of those of f on L_θ.

Proof. Since f is self-inversive it is certainly of type $(p, n - 2p, p)$ for some nonnegative integer p. By Theorem 4.4 this is equivalent to $\tilde{f}$ being of type $(p + k, n - 2p - k, p)$ for all sufficiently small $\xi \in L_{\theta+\pi/2}$ and left of L_θ.

By Lemma 4.3(a), we can apply Theorem 5.2 to $\tilde{f}$ to conclude that $\tilde{f}$ is of type $(p + k, n - 2p - k, p)$ if and only if $(\tilde{f})^\vee$ is of type $(p + k - 1, n - 2p - k, p)$

for w left of L_θ and of type $(p + k, n - 2p - k, p - 1)$ for w right of L_θ. The first part of the theorem follows then from Lemma 4.3(c) and the second part from the last statement in Theorem 4.4.

The significance of Theorems 5.2 and 5.3 is that they show that we can always reduce the question of finding the type of any polynomial of degree n to the same question for a polynomial of degree $n - 1$. In the next section we apply these theorems to establish algorithms for the classification of polynomials, which is a special case of the problem of finding the type.

6. Classification of polynomials. In this section we present necessary and sufficient conditions for a polynomial of degree n to belong to any of the classes defined in § 2. These conditions will be in terms of a polynomial of degree $n - 1$, and clearly by repeated application of our results an effective algorithm can be obtained. Our results are easy to apply in practice since the only expressions involved for a polynomial f are $f(w), f^*(w)$, $\check{f}$ and f', all of which are easily obtained.

THEOREM 6.1. *Suppose f is a polynomial of degree n and let w be any point left of L_θ. Then f is a Petrowsky–Gårding polynomial if and only if either $|f^*(w)| > |f(w)|$ and $\check{f}$ is Petrowsky–Gårding or $\check{f} \equiv 0$ and $nf(z) - (z - w^*)f'(z)$ is Petrowsky–Gårding.*

Proof. Suppose f is a Petrowsky–Gårding polynomial. Then f is of type $(p, n - p, 0)$ for some nonnegative integer p and

$$\left|\frac{f(w)}{f^*(w)}\right| = \prod_{j=1}^{n} \left|\frac{w - z_j}{w - z_j^*}\right|,$$

where the z_j are on or left of L_θ. But since also w is left of L_θ,

$$|w - z_j^*| = |w - \alpha| + |\alpha - z_j^*|,$$

where $\alpha \in L_\theta$ is the intersection of L_θ and the line joining w and z_j^*. Moreover, by the triangle inequality,

$$|w - z_j| \leqq |w - \alpha| + |\alpha - z_j|$$

with equality if and only if z_j (and hence also z_j^*) lies on L_θ. Noting that $|\alpha - z_j^*| = |\alpha - z_j|$ we conclude that $|w - z_j| \leqq |w - z_j^*|$ with equality if and only if z_j lies on L_θ. This implies at once that $|f^*(w)| \geqq |f(w)|$ with equality if and only if each z_j lies on L_θ. First we consider the case $|f^*(w)| > |f(w)|$. Then by Theorem 5.2, $\check{f}$ is of type $(p - 1, n - p, 0)$ and hence Petrowsky–Gårding. On the other hand, if $|f^*(w)| = |f(w)|$, then we have shown that f is conservative and therefore also self-inversive, so that $\check{f} \equiv 0$ and by Theorem 5.3, $nf(z) - (z - w^*)f'(z)$ is of type $(k - 1, n - k, 0)$, where k is the number of distinct zeros of f on L_θ, and thus it is Petrowsky–Gårding.

To establish the reverse implications suppose first that $|f^*(w)| > |f(w)|$ and $\check{f}$ is Petrowsky–Gårding. Then $\check{f}$ is of type $(p - 1, n - p, 0)$ for some positive integer p, so that by Theorem 5.2, f must be of type $(p, n - p, 0)$. On the other hand, if $\check{f} \equiv 0$, then by Corollary 3.2, f is of type $(p, n - 2p, p)$ for some nonnegative integer p. By Theorem 5.3, $nf(z) - (z - w^*)f'(z)$ is then of type $(p + k - 1, n - 2p - k, p)$, and by hypothesis it is Petrowsky–Gårding. These properties can

be reconciled only if $p = 0$, and we conclude that f is of type $(0, n, 0)$. In both cases therefore f is Petrowsky–Gårding.

The remaining classes are easily recognized by the following criteria. Since the proofs of these are merely special cases of the proof of Theorem 6.1 we state them as corollaries, and we leave the necessary modifications in their proofs to the reader.

COROLLARY 6.2. *Let w be any point left of L_θ. Then f is a Hurwitz polynomial if and only if $|f^*(w)| > |f(w)|$ and $\check{f}$ is Hurwitz.*

COROLLARY 6.3. *Suppose f is a polynomial of degree n and let w be any point left of L_θ. Then f is a conservative polynomial if and only if $\check{f} \equiv 0$ and $nf(z) - (z - w^*)f'(z)$ is Petrowsky–Gårding.*

COROLLARY 6.4. *Suppose f is a polynomial of degree n and let w be any point left of L_θ. Then f is a simple Petrowsky–Gårding polynomial if and only if either $|f^*(w)| > |f(w)|$ and $\check{f}$ is Petrowsky–Gårding or $\check{f} \equiv 0$ and $nf(z) - (z - w^*)f'(z)$ is Hurwitz.*

COROLLARY 6.5. *Suppose f is a polynomial of degree n and let w be any point left of L_θ. Then f is a simple conservative polynomial if and only if $\check{f} \equiv 0$ and $nf(z) - (z - w^*)f'(z)$ is Hurwitz.*

7. Examples. We apply our results now to some specific examples for $\theta = \pi/2$. For convenience we take $w = -1$. Our first example is

$$f(z) = z^5 + 2z^4 + 4z^3 + 4z^2 + 2z + 1$$

and we construct the sequence of polynomials

$$\begin{aligned}
f_0(z) &= f(z),\\
f_1(z) &= f_0(z)/(z+1) = z^4 + z^3 + 3z^2 + z + 1,\\
f_2(z) &= \check{f}_1(z)/2 = 2z^3 + 3z^2 + 3z + 2,\\
f_3(z) &= f_2(z)/(z+1) = 2z^2 + z + 2,\\
f_4(z) &= \check{f}_3(z)/4 = z + 1.
\end{aligned}$$

Determining the sign of $|f_j^*(-1)| - |f_j(-1)|$ for $j = 0, \cdots, 3$ and noting that f_4 is of type $(1, 0, 0)$, we conclude at once that f is of type $(5, 0, 0)$.

Consider now the polynomial

$$f(z) = 9z^5 + 4z^4 + 5z^3 + 4z^2 + 2z + 1.$$

This gives rise to the sequence

$$\begin{aligned}
f_0(z) &= f(z),\\
f_1(z) &= \check{f}_0(z)/2 = 81z^4 - 17z^3 + 62z^2 + 2z + 16,\\
f_2(z) &= -\check{f}_1(z)/6 = 405z^3 + 496z^2 - 186z + 80,\\
f_3(z) &= -\check{f}_2(z)/96 = 4860z^2 - 2597z + 365,\\
f_4(z) &= -\check{f}_3(z)/25970 = 972z + 73,
\end{aligned}$$

from which we see that f is of type $(3, 0, 2)$, since f_4 is of type $(1, 0, 0)$.

In [1] Bareiss considers the polynomial

$$f(z) = z^3 - 5z^2 + (9 + 4i)z - (1 + 8i).$$

For this we construct the sequence

$$f_0(z) = f(z),$$
$$f_1(z) = -\check{f}_0(z)/4 = (3 + 2i)z^2 + (22 + 18i)z + (21 - 20i),$$
$$f_2(z) = \check{f}_1(z)/204 = (1 + i)z + (1 - i),$$

which shows, on examining the signs of $|f_j^*(-1)| - |f_j(-1)|$ for $j = 0, 1, 2$ and noting that f_2 is of type $(0, 1, 0)$, that f is of type $(0, 1, 2)$. Furthermore, f_2 is the self-inversive factor of f, so that one zero of f is i and the others are then easily seen to be $2 + i$ and $3 - 2i$.

As a final example we consider the polynomial

$$f(z) = z^5 + (1 + 2i)z^4 + (-1 + 2i)z^3 + (1 - 2i)z^2 + (-2 + 2i)z - 4i.$$

For this the appropriate sequence is

$$f_0(z) = f(z),$$
$$f_1(z) = \check{f}_0(z)/(4 - 8i) = z^4 + (-2 + 2i)z^3 + (1 - 4i)z^2 + (-2 + 2i)z - 4i,$$
$$f_2(z) = -\check{f}_1(z)/(8 - 16i) = z^3 + 2iz^2 + z + 2i,$$
$$f_3(z) = 3f_2(z) - (z - 1)f_2'(z) = (3 + 2i)z^2 + (2 + 4i)z + (1 + 6i),$$
$$f_4(z) = \check{f}_3(z)/4 = (7 - 8i)z + (13 + 8i).$$

Since f_4 is of type $(1, 0, 0)$, $|f_3^*(-1)| > |f_3(-1)|$, f_2 is self-inversive, $|f_1^*(-1)| < |f_1(-1)|$ and $|f_0^*(-1)| < |f_0(-1)|$, we can conclude that f is of type $(1, 3, 1)$ and that the three zeros on the imaginary axis are distinct (the zeros of f are of course $1, -2, \pm i$ and $-2i$).

REFERENCES

[1] E. H. BAREISS, *The numerical solution of polynomial equations and the resultant procedures*, Mathematical Methods for Digital Computers, vol. II, A Ralston and H. S. Wilf, eds., John Wiley, New York, 1967, pp. 185–214.

[2] F. F. BONSALL AND MORRIS MARDEN, *Zeros of self inversive polynomials*, Proc. Amer. Math. Soc., 3 (1952), pp. 471–475.

[3] R. J. DUFFIN, *Algorithms for classical stability problems*, SIAM Rev., 11 (1969), pp. 196–213.

[4] E. FRANK, *On the real parts of the zeros of complex polynomials and applications to continued fraction expansions of analytic functions*, Trans. Amer. Math. Soc., 62 (1947), pp. 272–283.

[5] L. GÅRDING, *An inequality for hyperbolic polynomials*, J. Math. Mech., 8 (1959), pp. 957–966.

[6] A. HURWITZ, *Über die Bedingungen unter welchen eine Gleichung nur Wurzeln mit negativen reellen Teilen besitzt*, Math. Ann., 46 (1895), also Werke vol. 2, pp. 533–545.

[7] E. J. ROUTH, *The Advanced Part of a Treatise on the Dynamics of a System of Rigid Bodies*, Macmillan, London, 1905.

[8] J. H. WILKINSON, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N.J., 1963.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

FINITE-DIMENSIONAL APPROXIMATIONS OF STATE-CONSTRAINED CONTINUOUS OPTIMAL CONTROL PROBLEMS*

JANE CULLUM†

Abstract. "Continuous" optimal control problems with constraints on the control variables and on the state variables are considered. For any such problem, P, using ideas from Cullum [10] and Warga [15], a sequence of finite-dimensional optimization problems associated with P is constructed. It is proved (under certain hypotheses on P) that extremals (in particular, optimal solutions) of these finite-dimensional problems approximate, in a desirable sense, extremals of the original "continuous" problem P.

1. Introduction. For the purposes of this paper, a state-constrained continuous optimal control problem is a control problem of the following type: r-dimensional, vector-valued control (input) functions u and $(n+1)$-dimensional, vector-valued state (output) functions x are related by a system of ordinary differential equations,

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

These inputs and outputs are observed over a fixed time interval $[0, T]$. At each time t, any input $u(t)$ is restricted to a set $U \subset E^r$ (control constraint), and any output $x(t)$ is restricted to a set $A \subset E^{n+1}$ (intermediate state constraint). Moreover, $x = (x_1, \hat{x})$, $f = (f_1, \hat{f})$, $\hat{x}(0)$ must be in a set $G^0 \subset E^n$, and $\hat{x}(T)$ must be in a set $G^1 \subset E^n$ (initial and terminal state constraints). Determine the minimum value of the first coordinate of the output at the final time, $x_1(T)$; and obtain an input u that generates an output that attains this minimum. A problem of this type will be denoted by P, and its optimal cost will be denoted by $C(P)$. It is assumed that each problem is deterministic and has an optimal solution.

DEFINITION. Let P_m, $m = 1, 2, \cdots$, be a sequence of finite-dimensional optimization problems (discrete optimal control problems) such that for each m, P_m has the following form. Minimize the first coordinate of the vector $x_m^{q(m)}$ over all trajectories and controls $(\bar{x}_m, \bar{u}_m)$ with $\bar{x}_m = (x_m^0, \cdots, x_m^{q(m)})$ and $\bar{u}_m = (u_m^0, \cdots, u_m^{q(m)-1})$ such that for $k = 0, 1, \cdots, q(m) - 1$,

$$x_m^{k+1} = g_m^k(x_m^{k+1}, x_m^k, u_m^k),$$

$$u_m^k \in U_m \subset E^r, \quad x_m^k \in A_m \subset E^{n+1}, \tag{1.1}$$

$$x_m^0 \in G_m^0, \quad x_m^{q(m)} \in G_m^1.$$

If (i) for large m, pairs $(\bar{x}_m, \bar{u}_m)$ admissible for P_m exist; and (ii) there exists a continuous problem P such that the optimal costs of the problems P_m converge to the optimal cost of P as $m \to \infty$; and any sequence of optimal solutions of the P_m, $m = 1, 2, \cdots$, "converges" (in some meaningful sense) to an optimal solution of P; then P_m, $m = 1, 2, \cdots$, is a *finite-dimensional* or *discrete approximation to P*.

* Received by the editors January 15, 1971, and in revised form January 27, 1972.

† Thomas J. Watson Research Center, IBM Corporation, Yorktown Heights, New York 10598.

In this paper an attempt is made to characterize those state-constrained continuous optimal control problems that possess discrete approximations, and to determine an explicit procedure for constructing such an approximation. The objective of this procedure is to solve P (approximately) by solving one or more of the finite-dimensional problems in the approximation.

This work is motivated by the following. (i) State-constrained problems, for example, [1], [2], [3], [4], [5], have been solved by discretizing the time, replacing the system of differential equations by difference equations and solving the resulting discrete optimal control problems. It seems desirable to establish the legitimacy of this procedure.

(ii) Attempts to utilize extensions of Pontryagin's maximum principle to solve state-constrained problems have been successful only in special cases because of the disparate treatment of the control and the state constraints, and the discontinuous multiplier functions. For example, see [6], [7]. In [6] there are no control or terminal state constraints, $U = E^r$ and $G^1 = E^n$. In [7], the control is an unconstrained scalar, and the optimal trajectory consists of three sections, two interior to the state constraint and one boundary arc that can be solved for separately. If a state-constrained problem is approximated by discrete optimal control problems, then each of these problems can be considered as a mathematical programming problem [8], the constraints on the inputs and outputs can be treated directly (assuming, for example, that these quantities must satisfy simple boundedness conditions) and it is not necessary to introduce a maximum principle in which the control and the state play different roles.

(iii) Any method for solving a continuous optimal control problem requires that various approximations and/or discretizations be made to obtain numerical results. Thus, the question whether to apply the maximum principle and then to discretize (replace the differential equations in the necessary conditions for optimality by difference equations), or to discretize the problem initially and then solve the resulting finite-dimensional or discrete optimization problem arises naturally.

If a continuous problem P is well-posed; that is, it has a unique solution and this solution depends "continuously" upon all the data in P, then it seems clear that any discretization of P that is "consistent" and for which admissible pairs exist will provide a discrete approximation to P. However, most optimal control problems are not well-posed and except for special cases it is not possible to determine a priori whether or not a given problem is well-posed. Hence, it is desirable to obtain results that apply to a more general class of optimal control problems.

The following discussion demonstrates the inadvisability of blindly discretizing a given continuous optimal control problem.

Example 1. Minimize the time to transfer the point $x^0 = (0.875, 2.5)$ to the point $x^f = (2.0, 2.0)$ along trajectories of the system

$$\dot{x}_1(t) = x_2(t), \quad \dot{x}_2(t) = u(t), \qquad t \in [0, T],$$
$$-1 \leqq u(t) \leqq 1, \quad A = E^2.$$

Denote this problem by P^1. P^1 is completely controllable. By inspection, the

optimal solution is: optimal time, $t^* = 0.5$, optimal control, $u^*(t) \equiv -1$, and the points x^0 and x^f lie on the parabola $2x_1^*(t) = -(x_2^*(t))^2 + 8$. An obvious discretization of P^1 is the following. Pick a $(\Delta t)_m$, for example, $1/m$, $m = 1, 2, \cdots$. Replace the differential equations by difference equations (for example, use the Cauchy–Euler explicit difference scheme). Since the system is linear, for any k, $k = 1, 2, \cdots$,

$$x_1^k = \sum_{j=0}^{k-1} \frac{x_2^j}{m} + 0.875,$$
$$x_2^k = \sum_{j=0}^{k-1} \frac{u^j}{m} + 2.5. \tag{1.2}$$

For each m, define P_m as follows. Minimize k such that $x_1^k = 2.0$, $x_2^k = 2.0$, and $-1 \leqq u^j \leqq 1$, $j = 0, \cdots, k-1$. That is, for each m determine the minimum number of time periods required to reach the desired terminal point.

For fixed m, the points reachable for $k = 1, 2, \cdots$ are contained in the rectangle with center $(0.875 + 2.5k/m, 2.5)$, width $k(k-1)/m^2$, and height $2k/m$. The first rectangle that intersects the line $x_2 = 2.0$ corresponds to $k = 0.5m$ (assume m is even). It is easy to prove that the x_1-coordinate of any admissible point in the rectangle for $k = 0.5m$ is bigger than 2.0. That is, x_2^* is monotone decreasing and $u^*(t) \equiv -1$. Hence, for any admissible control, $x_2^*(t) < x_2^j$ on each interval $[j/m, j + 1/m]$. Therefore,

$$x_1^{0.5m} = \sum_{j=0}^{0.5m-1} \frac{x_2^j}{m} + 0.875 > \int_0^{0.5} x_2^*(t)\,dt + 0.875 = 2.0.$$

Observe further that for $k < 2.5m$, $x_2^k > 0$. Therefore, for $k \leqq 2.5m$, $x_1^k > 2$, and the point (2.0, 2.0) cannot be in any rectangle corresponding to a $k < 2.5m$. In fact, this point is not reachable for any $k < 4.5m$, $m = 1, 2, \cdots$. Two units of time is a reasonable estimate on the length of time required to transfer x^0 to x^f. Hence, four units of time or $4m - 1$ subintervals seems a reasonable guess for P_m^1, $m = 1, 2, \cdots$. However, if only these values of k are considered, then none of the P_m^1, $m = 1, 2, \cdots$, have admissible pairs. On the other hand, if k is not constrained, then admissibility or feasibility is attained for some $k > 4.5m$. Hence, the optimal cost of P_m^1, $C(P_m^1)$, is bigger than $4.5m$. Since $C(P^1) = 0.5$, it is clear that these costs cannot converge to the optimal cost of P^1.

Remarks. If $x^f = (0, 0)$ in P^1, then the obvious discretization does produce a discrete approximation (see [9]). In this case, P^1 is well-posed. The simple explicit difference scheme was used in Example 1 and elsewhere in this paper only because it is the simplest scheme to write down. Any other convergent difference scheme, explicit or implicit in x, could be used (see [10]).

Example 1 demonstrates that the question: Given a problem P, how does one determine a discrete approximation to P, is a real question. The preceding discussion does not imply that no discrete approximation of P^1 exists. It only demonstrates that the obvious approach does not produce a discrete approximation to P^1. In fact, the construction given in [10] can be used to obtain an appropriate discretization. This construction requires knowledge of an interval $[0, T]$ such that an admissible solution of P^1 exists on some interval $[0, T^*) \subset [0, T]$.

Set $t = t_f s$ and transform P^1 into the following fixed time problem. Minimize t_f over all pairs $((t_f, y), v)$ such that

$$\begin{aligned} \dot{y}_1(s) &= y_2(s)t_f(s), \\ \dot{y}_2(s) &= v(s)t_f(s), \\ \dot{t}_f(s) &= 0, \quad 0 \leqq s \leqq 1, \quad -1 \leqq v(s) \leqq 1, \\ (t_f(0), y(0)) &\in G^0 = \{0 \leqq t_f(0) \leqq 2, y(0) = (0.875, 2.5)\}, \\ (t_f(1), y(1)) &\in G^1 = \{t_f \text{ free}, y(1) = (2.0, 2.0)\} \end{aligned} \tag{1.3}$$

(observe that T was set equal to 2). The differential equations are linear in the control v and satisfy the other hypotheses of [10]. Two steps are required.

Step 1 (*Feasibility*). Define $B_m, m = 1, 2, \cdots$, as follows. Minimize ρ over all $((t_f^m, \bar{y}_m), \bar{v}_m)$ with

$$\begin{aligned} y_{1\,m}^{k+1} - y_{1m}^k &= y_{2m}^k t_f^k/m, \\ y_{2m}^{k+1} - y_{2m}^k &= v_m^k t_f^k/m, \\ t_f^{k+1} - t_f^k &= 0, \quad -1 \leqq v_m^k \leqq 1, \quad k = 0, 1, \cdots, m-1, \\ (t_f^0, y_m^0) &\in G^0 = \{0 \leqq t_f^0 \leqq 2, y_m^0 = (0.875, 2.5)\}, \\ (t_f^m, y_m^m) &\in G_m^1 = \{t_f^m \text{ free}, y_m^m \in S((2.0, 2.0), \rho)\}, \end{aligned} \tag{1.4}$$

where $S(z, r)$ denotes the sphere of radius r with center z.

Step 2 (*Optimization*). Define P_m, $m = 1, 2, \cdots$, as follows. For each m, minimize t_f^m over all $((t_f^m, \bar{y}_m), \bar{v}_m)$ satisfying equations (1.4) with $\rho = C(B_m)$. For P^1, the maximum principle is sufficient. Hence, by [10], the optimal cost of P_m converges to the optimal cost of P as $m \to \infty$, and optimal solutions of the P_m converge to the optimal solution of P as $m \to \infty$. The trajectories converge uniformly, and the controls converge in the L_2-topology.

2. Historical development. Neustadt in [11] considered the discretization of a particular linear, time optimal control problem. He constructed a family of discrete optimal control problems by discretizing the solution formula for the associated differential equations. He used the fact that the optimal controls of the discrete and of the continuous problems could be obtained explicitly to prove that the optimal costs and the optimal controls of the discrete problems converge to the corresponding optimal cost and control of the continuous problem as the discretization is refined.

The techniques in [11] and [12] were used in [9] to prove that any controllable linear problem with a convex cost functional and convex constraint sets A, U, G^0, and G^1 possesses a discrete approximation. The proof, however, was not constructive, other than demonstrating that such an approximation can be obtained by properly enlarging the constraint sets A and U; no formula for choosing these enlargements was given. A similar result was obtained for time optimal control problems when the terminal states satisfy an additional hypothesis. The proof for time optimal control problems is constructive when $A = E^{n+1}$, and essentially states that the obvious approach yields a discrete approximation.

Budak et al. [13] considered continuous, fixed time problems without any

state constraints, terminal or intermediate, i.e., $G^1 = E^n$, and $A = E^{n+1}$, and proved that the optimal costs of the discrete optimization problems obtained by directly discretizing the original problem in almost any fashion converge to the optimal cost of the original problem. If no state constraints are present, the question of the existence of feasible pairs for the discrete problems evaporates, so this is an easy case.

In [14] Daniel considered a general state-constrained problem and the discrete optimal control problems generated from it by discretizing the differential equations and enlarging the constraint sets, and attempted to identify the properties necessary to guarantee convergence of the optimal solutions of the discrete problems to optimal solutions of the continuous problem as the enlargements in the control and state constraint sets, and the sample time, Δt, are reduced to zero. His results depend upon the existence of maps with certain properties.

The results in [10] are constructive. Fixed time problems whose system equations are linear in the control, whose cost functionals are convex in u, and for which $A = E^{n+1}$ are considered. No special hypotheses like controllability are required. Whenever P is a problem for which Pontryagin's maximum principle is a sufficient condition of optimality, the procedure in [10] yields a discrete approximation to P. In general, however, one can state only that limit points of any sequence of optimal solutions of the discrete optimal control problems exist, and any such limit point is an extremal of P. This type of result is typical in optimization theory where the necessary conditions of optimality play a vital role in the proofs of theorems. For example, with respect to the algorithm of steepest descent which is used to determine the minimal point of a function of n variables; if the function being minimized, f, is strictly convex, then the iterates generated by this algorithm converge to the minimal point of f. In general, however, one can state only that any limit point of this sequence of iterates is a stationary point of f [26]. The implication here is that a given set of iterates could oscillate between several limit points and not display any discernible convergence. In practice, however, this type of behavior rarely occurs and the iterates generated usually converge to the closest minimal point of f.

This paper extends the results in [10] to state-constrained problems; that is, $A \neq E^{n+1}$. A general state-constrained continuous optimal control problem P is considered and a sequence of finite-dimensional optimization problems (mathematical programming problems) associated with P is constructed. Whenever P satisfies certain convexity hypotheses, and the extended maximum principle in Warga [15] is a sufficient condition for optimality for P, the sequence constructed is a discrete approximation to P. If these conditions are only necessary, then as in [10], one can state only that any limit point of any sequence of extremals (in particular optimal solutions) of the finite-dimensional problems constructed is an extremal of P. The constructive procedure presented is used in § 8 to solve an example problem numerically.

3. Construction of the approximations. A global assumption throughout the paper is:

(A1) $A = \{x | a^k(\hat{x}) \leqq 0, k = 1, \cdots, q\}$, where each a^k is a convex, continuously differentiable function, and the interior of A is nonempty.

For simplicity of notation assume that $q = 1$ and set $a^1 = a$. The changes needed to handle $q > 1$ are not fundamental.

Using the results from [10], one is led to propose the following discretizations. Partition the time interval, for example, choose $\Delta t_m = T/m$. The determination of each member in the discrete approximation requires two steps. The first step (feasibility) yields a set of equalities and inequalities that have solutions. The second step (optimality) determines a solution of this set of equalities and inequalities that minimizes a particular discrete analogue of the original cost function.

Step 1 (*Feasibility*). Let B_m^*, $m = 1, 2, \cdots$, denote the following discrete optimal control problem: Minimize the scalar ρ over all pairs $(\bar{x}_m, \bar{u}_m) = (x^0, x^1, \cdots, x^m, u^0, \cdots, u^{m-1})$, where $x^j \in E^{n+1}$ and $u^j \in E^r$ such that

$$x^{k+1} - x^k = f(x^k, u^k)\Delta t, \qquad k = 0, 1, \cdots, m-1, \tag{3.1}$$

$$\hat{x}^0 \in G^0, \quad u^k \in U, \qquad k = 0, 1, \cdots, m-1, \tag{3.2}$$

$$\hat{x}^m \in \cup(G^1, \rho), \quad a(x^k) \leqq \rho, \quad \rho \geqq 0, \qquad k = 0, 1, \cdots, m. \tag{3.3}$$

The subscript m on Δt, x^k, and u^k has been dropped for notational simplicity. $\cup(G^1, \rho) = \{\hat{x} | d(\hat{x}, G^1) \leqq \rho\}$, where $d(\hat{x}, G^1)$ is the infimum of the Euclidean distance of $\hat{x}$ to any element $y \in G^1$.

In [10] with $A = E^{n+1}$ an appropriate set of equalities and inequalities (feasibility) was obtained by determining the minimum enlargement in the terminal constraint set G^1 necessary to obtain quantities that satisfy both the control constraints and the difference equations. With $A \neq E^{n+1}$, an enlargement or relaxation of the intermediate state constraint set A has also been included in (3.3). Let $\rho_m = C(B_m^*)$, the optimal cost of B_m^*.

Step 2 (*Optimization*). Let P_m^*, $m = 1, 2, \cdots$, denote the following discrete optimal control problem: Minimize the first component of x^m over all $(\bar{x}_m, \bar{u}_m)$ satisfying (3.1), (3.2), and (3.3) with $\rho = \rho_m$.

The proof in [10] that the sequence of discrete optimal control problems constructed is a discrete approximation to P (for a certain class of problems) was obtained by demonstrating concrete relationships between the necessary conditions of optimality associated with the discrete problems and the necessary conditions of optimality associated with the continuous problem. One would like to utilize that technique here. However, as will be demonstrated in § 7, the NCO (necessary conditions of optimality) associated with the particular problems P_m^*, $m = 1, 2, \cdots$, are only weakly related to the NCO associated with the original continuous problem. It is only possible to argue heuristically, not rigorously, that extremals (in particular, optimal solutions) of these problems converge to extremals (optimal solutions) of P as $m \to \infty$. To obtain a rigorous proof, the discretization must be altered. The idea for the alteration is taken from Warga [15]. In [15] Warga derives a set of necessary conditions of optimality for a state-constrained optimal control problem: by first, constructing a sequence Q_m, $m = 1, 2, \cdots$, of continuous optimal control problems with $A = E^{n+m+1}$ that "approximate" P; second, writing down the necessary conditions of optimality associated with each Q_m; and third, proving rigorously that limiting arguments are valid for the

corresponding sequence of multipliers and conditions generated, and hence, can be used to obtain necessary conditions of optimality for the limiting (original) problem P. A similar approach will be used here.

Replace B_m^* and P_m^* by the following.

Step 1 (*Feasibility*). Let B_m, $m = 1, 2, \cdots$, denote the following problem: Minimize the scalar ρ over all triplets $(\bar{x}_m, \bar{u}_m, \bar{v}_m) = (x^0, x^1, \cdots, x^m, u^0, \cdots, u^{m-1}, v^0, \cdots, v^m)$, where $x^j \in E^{n+1}$, $u^j \in E^r$, $v^j \in E^1$ such that

$$x^{k+1} - x^k = f(x^k, u^k)\,\Delta t, \qquad k = 0, 1, \cdots, m-1, \tag{3.1}$$

$$\hat{x}^0 \in G^0, \quad u^k \in U, \qquad k = 0, 1, \cdots, m-1, \tag{3.2}$$

$$v^{k+1} - v^k = b(x^k, u^k)\,\Delta t, \qquad k = 0, 1, \cdots, m-1, \tag{3.4}$$

$$\hat{x}^m \in \bigcup(G^1, \rho), \quad v^0 = a(\hat{x}^0), \tag{3.5}$$

$$v^k \leqq \rho, \quad \rho \geqq 0, \qquad k = 1, \cdots, m, \tag{3.6}$$

where $b(x, u) = \langle a_x(x), f(x, u)\rangle$; and $\langle c, d\rangle$ denotes the inner product of two vectors c and d in the related Euclidean space. For simplicity, $q = 1$ and the subscripts m were deleted. Observe that (3.1) and (3.2) are unchanged. The constraint function along a trajectory x, $a(x(t))$, has been entered as a new variable v that is constrained at each k. The constraints on v can be converted into terminal constraints as follows [15]. Let X_j denote the characteristic function of the interval $[0, t_j]$, where $t_j = jT/m$. Let $v_{jm}, j = 1, \cdots, m$, denote the solutions of the following difference equations:

$$v_{jm}^{k+1} - v_{jm}^k = b(x^k, u^k)\,\Delta t X_j(t^k), \qquad k = 1, \cdots, m. \tag{3.7}$$

Then for each m, B_m has the following $(m + n + 1)$-dimensional formulation with no intermediate state constraints ($A = E^{n+m+1}$): Minimize ρ over all triplets $(\bar{x}_m, \bar{u}_m, \bar{v}_m)$, where $\bar{v}_m = (v_m^0, \cdots, v_m^m)$, $v_m^k = (v_{1m}^k, \cdots, v_{mm}^k)$, satisfying (3.1), (3.2), (3.5), (3.7) and

$$v_{jm}^0 = a(\hat{x}^0), \quad v_{jm}^m \leqq \rho, \quad \rho \geqq 0, \qquad j = 1, \cdots, m. \tag{3.8}$$

Step 2 (*Optimality*). For each m, $m = 1, 2, \cdots$, define P_m as follows: Minimize the first component of x^m over all $(\bar{x}_m, \bar{u}_m, \bar{v}_m)$ satisfying (3.1), (3.2), (3.5), (3.7) and (3.8) with $\rho = \rho_m \equiv C(B_m)$, the optimal cost of B_m.

Remarks. Each P_m is a particular discretization of a particular continuous problem Q_m considered by Warga [15]. If $q > 1$, then q sets of equations like (3.7) are required. In actual computational work the original formulation of B_m would be utilized. Nonnegativity constraints such as (3.6) are natural constraints in mathematical programming problems. The key to the formulation of B_m (P_m) is the reduction of the intermediate state constraint to one involving the control (actually the state and the control).

The sample size $\Delta t = T/m$ and the explicit difference scheme were chosen only for notational convenience. Obviously, more sophisticated schemes would provide better approximations. $\bar{x}_m$ is called a trajectory of B_m (P_m), and $\bar{u}_m$ is similarly called a control function. As indicated by the notation used previously, $\bar{x}_m(\bar{u}_m)(\bar{v}_m)$ denotes the mapping of the time sequences $\{0, t_m^1, \cdots, T\}$

$(\{0, t_m^1, \cdots, t_m^{m-1}\})$ $(\{0, t_m^1, \cdots, T\})$ onto the sequence $\{x^0, x^1, \cdots, x^m\}$ $(\{u^0, \cdots, u^{m-1}\})$ $(\{v_m^0, \cdots, v_m^m\})$. The hat on x will usually be omitted when it is clear which x is being considered.

4. Necessary conditions of optimality. The verification that the sequence P_m, $m = 1, 2, \cdots$, is a discrete approximation to P (for a certain subclass of problems) relates the associated necessary conditions of optimality (NCO). Several formulations of NCO for state-constrained continuous optimal control problems exist (Gamkrelidze [16], Hestenes [17], Neustadt [18], Chang [19]). The formulation given by Warga [15] was used because some of the constructions used in [15] are repeated here. Warga's derivation requires that certain sets associated with P be convex. Apparently, hypotheses of this nature are in fact necessary to obtain strong forms of the NCO (see Neustadt [18]). Examples exist [20] that demonstrate that for discrete optimal control problems hypotheses of convexity are necessary to obtain a maximum principle even when $A = E^{n+1}$. (This, of course, is not true in the continuous case.)

The first rigorous version of a discrete maximum principle was given by Halkin [20] for "convex" problems. Holtzman and Halkin [21] extended this result to "directionally convex" problems. Canon, Cullum and Polak [8, p. 91], using a general variational theory for finite-dimensional optimization problems, extended these results further and obtained a principle that applies to a wider variety of "directionally convex" problems (i.e., intermediate state constraints and initial and terminal sets described by sets of equations and inequalities). The main result in this paper, Theorem 5.1, does not require the complete generality of this result since the approximating problems P_m, $m = 1, 2, \cdots$, do not have intermediate state constraints. The generality of the results in [8, p. 91] will be utilized in a heuristic discussion in § 7.

The sets of necessary conditions of optimality derived in [15] and [8, p. 91] were derived under specific sets of hypotheses. Since these NCO are to be used in the proofs, it is necessary that P satisfy all of these hypotheses. Most of these hypotheses are standard conditions on the continuity, differentiability, and growth of the functions f and a; and convexity and compactness hypotheses on the constraint sets for u and x. The only hypotheses that will be explicitly stated in this paper are the nonstandard ones. First, there is a convexity hypothesis:

(A2) For each $x \in E^{n+1}$, $\alpha \in E^{n+1}$, the set $F(x, \alpha) = \{(y, \bar{y}) | y = f(x, u), \bar{y} = f_x^T(x, u)\alpha, u \in U\}$ is compact and β-directionally convex for $\beta = (-1, 0, \cdots, 0)$.

A set S is β-directionally convex if and only if each element in the convex hull of S is the sum of a vector in S and $\gamma\beta$ for some $\gamma \leqq 0$. The discrete maximum principle is applicable to the P_m^*, $m = 1, 2, \cdots$, only if the sets $f(x, U) = \{y | y = f(x, u), u \in U\}$ are β-directionally convex. A modified form of (A2) is used in Theorem 5.1: (A2)′ Each set $F(x, \alpha)$ is compact and convex.

In [15] Warga required that the sets $F(x, \alpha)$ be convex. He used this requirement to prove that each of the continuous, approximating problems Q_m, $m = 1, 2, \cdots$, that he constructed has an optimal solution, and that limits of the corresponding sequences of multipliers exist and satisfy the appropriate differential equations. It is easy to prove that this convexity assumption can be replaced by β-directional convexity.

Three further assumptions will be made. (A3) was used in [8, p. 91] in the derivation of the discrete maximum principle. (A3), (A4), and (A5) together guarantee nondegeneracy of the maximum principle given in [15]. As demonstrated by the computational results for Example 2, (A4) is not a necessary condition.

(A3) f is not a function of x_1.

(A4) There exists an $\alpha < 0$ such that if $a^k(x) = 0$, $k \in N \subseteq \{1, \cdots, q\}$, $a^k(x) < 0$, $k \notin N$, then $\inf_{u \in U} \sum_{k \in N} \mu^k \langle a_x^k(x), f(x, u) \rangle \leqq \alpha$ for all $\mu^k \geqq 0, \sum_{k \in N} \mu^k = 1$.

(A5) For each $\hat{x}$, the gradients of the inequality and equality constraints on $\hat{x}$ (other than the differential equation or difference equation constraints) are linearly independent. If the constraint functions are linear, then (A5) can be removed [8].

(A4), which has been written for any q, is a "returnability" condition. It is used in [19] and [15], and guarantees that there are no points on the boundary of A at which the only way to satisfy the constraints of the continuous problem is to move along the boundary of A. In some problems the expression in (A4), $(da(x(t))/dt$, if $q = 1)$, is not even a function of u (see Example 2 in § 8).

Given a continuous problem P, we say that $P \in \mathscr{A}$ if: (i) P satisfies assumptions (A1) through (A5); (ii) P satisfies the hypotheses in Warga [15] and Canon, Cullum and Polak [8, p. 86] on the differentiability, continuity, and growth of the functions f and a, and on the compactness and convexity of the sets U, A, G^0, and G^1. We say that $P \in \mathscr{A}'$ if (A2)′ replaces (A2).

The results of applying the maximum principle in [15] to P, and the principle in [8, p. 91] to P_m, $m = 1, 2, \cdots$, are stated in Theorems 4.1 and 4.2, respectively.

THEOREM 4.1 [15, pp. 452–453]. *Let $P \in \mathscr{A}$. Let x^* be any optimal solution of P. Let $C^1 \subset G^1$ be any compact, convex subset of G^1 that contains $\hat{x}^*(T)$. Let $Z = \{t | a(x^*(t)) = 0\}$. Then there exist multiplier functions $\mu : I \to E^1$, $z : I \to E^{n+1}$ and a control function $u^* : I \to E^r$ such that:*

$$\mu(t) \leqq 0 \quad \text{and} \quad |z(t)| + |\mu(t)| > 0, \qquad t \in I, \tag{4.1}$$

where $|z| = \sum_{j=1}^{n+1} |z_j|$;

$$\begin{aligned} &z \text{ is absolutely continuous on } I, \\ &\mu \text{ is nondecreasing on } I, \ \mu \text{ is constant on every subinterval} \\ &\text{of } I - Z, \text{ and } \mu(T) a(x^*(T)) = 0; \end{aligned} \tag{4.2}$$

$$\begin{aligned} &\dot{x}^*(t) = f(x^*(t), u^*(t)) \quad \text{and} \\ &\dot{z}(t) = -f_x^T(x^*(t), u^*(t)) z(t) - \mu(t) b_x(x^*(t), u^*(t)) \quad \text{a.e. in } I, \end{aligned} \tag{4.3}$$

where $b(x, u) = \langle a_x(x); f(x, u) \rangle$.

Set $v = z + \mu a_x(x^*)$; *then*

$$\langle v(t), f(x^*(t), u^*(t)) \rangle = \max_{u \in U} \langle v(t), f(x^*(t), u) \rangle \quad \text{a.e. in } I. \tag{4.4}$$

$$\begin{aligned} &\text{The maximum of } \langle \hat{v}(0), \hat{x} \rangle \text{ over } G^0 \text{ occurs at } \hat{x}^*(0). \\ &\text{The minimum of } \langle \hat{z}(T), \hat{x} \rangle \text{ over } C^1 \text{ occurs at } \hat{x}^*(T). \end{aligned} \tag{4.5}$$

There exists a $0 < t^* < T$ *such that* $|v(t)| \neq 0$ $(t^* \leqq t \leqq T)$ *and either*

(4.6)
(a) $t^* = 0$ *or*
(b) $t^* \in Z$, *and* $z(t^*) = -\gamma a_x(x^*(t^*))$ *for some* $\gamma \leqq \mu(t^*)$.

Theorem 4.2 applies to problems P with $G^0 = \{\hat{x} | g^0(\hat{x}) = 0\}$ and $G^1 = \{\hat{x} | q^1(\hat{x}) \leqq 0\}$. The more general case is considered in Corollary 5.4 to Theorem 5.1. Let $P \in \mathscr{A}$ and let P_m, $m = 1, 2, \cdots$, be the sequence of discrete optimal control problems defined by (3.1), (3.2), (3.5), (3.7) and (3.8). Let $g^0 : E^n \to E^a$, and $q^1 : E^n \to E^b$. Fix m and set $A^k = [\partial f / \partial u(x^k, u^k)]$, $b_x^k = b_x(x^k, u^k)$ and $a_x^k = a_x(x^k)$ (the subscripts m have been omitted). An application of the results in [8, p. 91] to P_m yields the following theorem.

THEOREM 4.2. *Let* $(\bar{x}_m, \bar{u}_m, \bar{v}_m)$ *be an optimal solution of* P_m. *Then there exist costate vectors* $(p^0, p^1, \cdots, p^m)$, $p^j \in E^{n+m+1}$, $j = 0, \cdots, m$, *a multiplier* $\beta \in E^{m+a}$, *and a multiplier* $\lambda \in E^{m+b}$, $\lambda_j \leqq 0$, $j = 1, \cdots, m + b$, *such that if* $p^j = (y^j, w^j)$, *where* $y^j \in E^{n+1}$ *and* $w^j \in E^m$, *then*

$$y_1^j = y_1^0 = \text{const.} \leqq 0, \qquad j = 1, \cdots, m, \tag{4.7}$$

$$|p^0| + \cdots + |p^m| + |\beta| \neq 0, \tag{4.8}$$

$$y^k - y^{k+1} = A^k y^{k+1} \Delta t + \sum_{s=k+1}^{m} w_s^{k+1} b_x^k \Delta t, \tag{4.9}$$

$$w^k - w^{k+1} = 0, \qquad k = 0, 1, \cdots, m-1.$$

The transversality conditions at $t = 0$ *are*

$$\begin{aligned} \hat{y}^0 &= -\sum_{i=1}^{a} (g_i^0)_x \beta_i + a_x\left(\sum_{i=a+1}^{m+a} \beta_i\right), \\ w_j^0 &= -\beta_{j+a}, \qquad j = 1, \cdots, m. \end{aligned} \tag{4.10}$$

The transversality conditions at $t = T$ *are*

$$\begin{aligned} \hat{y}^m &= \sum_{j=1}^{b} \frac{\partial q_j^1}{\partial x} \lambda_j, \\ w_j^m &= \lambda_{j+b}, \qquad j = 1, \cdots, m, \end{aligned} \tag{4.11}$$

$$\begin{aligned} \lambda_j (q_j^1(x^m) - \rho_m) &= 0, \qquad j = 1, \cdots, b, \\ \lambda_{j+b}(v_{jm}^m - \rho_m) &= 0, \qquad j = 1, \cdots, m. \end{aligned} \tag{4.12}$$

For $k = 0, 1, \cdots, m-1$, *the maximum over* $u \in U$ of

$$\left\langle y^{k+1} + \left(\sum_{j=k+1}^{m} w_j^{k+1}\right) a_x^k, f(x^k, u) \right\rangle \tag{4.13}$$

occurs at u^k.

The hats on x have been dropped. λ does not appear in (4.8) because it is not obtained as a direct result of the main theorem in [8, p. 27]. It is obtained from

an additional application of Farkas' lemma. Define $\bar{\mu}_m = (\mu_m^0, \cdots, \mu_m^m)$ as follows:

$$\mu_m^k = \sum_{s=k+1}^{m} w_s^{k+1}, \qquad k = 0, 1, \cdots, m-1,$$

and set $\mu_m^m = \mu_m^{m-1}$. Let μ_m denote the piecewise linear extension of $\bar{\mu}_m$ to I. Since $\lambda_j \leqq 0$, $j = 1, \cdots, m+b$, and (4.11) holds, μ_m is a nonpositive, nondecreasing function of t. The analogies between (4.9) and (4.13) in the discrete case, and (4.3) and (4.4) in the continuous case are obvious. The notation a_x^k has been used to denote two different quantities; in (A4) and in Theorem 4.2. However, since $q = 1$ and $a^1 = a$, there should not be any confusion.

5. Convergence of the approximations.

THEOREM 5.1. *Let $P \in \mathscr{A}'$. Assume that $G^0 = \{x|g^0(\hat{x}) = 0\}$ and $G^1 = \{x|q^1(\hat{x}) \leqq 0\}$. Let $\bar{x}_m$ be an extremal of the corresponding problem P_m, $m = 1, 2, \cdots$. Any limit point x^* of this sequence (in the uniform topology) is an extremal of P.*

Remarks. A trajectory $\bar{x}_m$ of P_m is an extremal of P_m if and only if there exist quantities $\bar{u}_m, \bar{v}_m, \bar{p}_m, \lambda_m, \beta_m$ such that (4.7) through (4.13) are satisfied. A trajectory x^* of P is an extremal of P if and only if there exist quantities u^*, μ^*, and z^* satisfying (4.1) through (4.6). The convergence of $\bar{x}_m$ to x^* refers to the convergence of the piecewise linear extensions of $\bar{x}_m$ to the interval $[0, T]$. As stated earlier, Theorem 5.1 is typical of the results in optimization theory; without knowledge that the continuous problem has a unique solution and that the necessary conditions of optimality (NCO) are in fact sufficient, one cannot assert that the sequence of iterates generated by a given scheme converges; or that if it converges, then it converges to the desired minimal point.

In each of the corollaries, P is assumed to satisfy the hypotheses of Theorem 5.1.

COROLLARY 5.1. *If the NCO for P are in fact sufficient, then the optimal cost of P_m converges to the optimal cost of P as $m \to \infty$. Moreover, any limit point of a sequence of optimal solutions (extremals) (near optimal solutions) of the P_m, $m = 1, 2, \cdots$, is an optimal solution of P. If, in addition, P has a unique optimal solution x^*, then, in fact, any such sequence converges to x^*; and in fact P_m, $m = 1, 2, \cdots$, is a discrete approximation to P.*

Corollary 5.1 identifies a class of state-constrained optimal control problems that have discrete approximations. The next corollary is important computationally. It states that enlargements of the terminal and intermediate state constraint sets that are larger than necessary for obtaining feasibility may be used.

COROLLARY 5.2. *Let $\tilde{\rho}_m \geqq C(B_m)$ and $\tilde{\rho}_m \to 0$ as $m \to \infty$. Let $\tilde{P}_m$, $m = 1, 2, \cdots$, be obtained from P_m by replacing the cost $C(B_m)$ by $\tilde{\rho}_m$. Then the conclusions of Theorem 5.1 are valid for $\tilde{P}_m$, $m = 1, 2, \cdots$.*

Corollary 5.3 states that the relaxation in the terminal state constraints that was used (along with a relaxation in the intermediate state constraints) to get feasibility may be incorporated into the difference equations. As stated, each problem B_m, $m = 1, 2, \cdots$, has a quadratic constraint, namely, $\hat{x}_m^m \in \cup(G^1, \rho)$. If the constraints in the original problem are linear, then the introduction of this nonlinearity is very undesirable. Corollary 5.3 states that the desired linearity can be preserved.

COROLLARY 5.3. *Let* D_m, $m = 1, 2, \cdots$, *be the discrete optimal control problems obtained from* B_m, $m = 1, 2, \cdots$, *by inserting a vector* $\rho \in E^n$ *into the last* n *of the* $n + 1$ *difference equations obtained in* (3.1) *when* $k = m - 1$; *removing the scalar* ρ *from the terminal constraint set*; *replacing the* ρ *in the constraints on* v *by* $|\rho|$, *the* L_1*-norm of* ρ; *and then minimizing* $|\rho|$. *Let* $\tilde{P}_m$, $m = 1, 2, \cdots$, *be the problems corresponding to* P_m *obtained from* D_m. *Then the conclusions of Theorem* 5.1 *and its corollaries are valid for the* $\tilde{P}_m$, $m = 1, 2, \cdots$.

Finally, consider Corollary 5.4 which states that Theorem 5.1 is valid for more general initial and terminal state constraint sets.

COROLLARY 5.4. *Let* $G^0 = \{\hat{x} | g^0(\hat{x}) = 0, q^0(\hat{x}) \leqq 0\}$ *and* $G^1 = \{\hat{x} | q^1(\hat{x}) \leqq 0\}$, *where* $g^0 : E^n \to E^a$, $q^0 : E^n \to E^c$ *and* $q^1 : E^n \to E^b$, *and the Jacobian of* (g^0, q^0) *has full rank. Then the conclusions of Theorem* 5.1 *are valid.*

Theorem 5.1 is of course applicable when $A = E^{n+1}$. Hence, it generalizes the results in [10] to more general convex problems. The proof in [10] relied heavily upon the linearity of f in the control u. In this paper the linearity in u has been replaced by assumption (A2)′. Two lemmas are needed in the proof.

LEMMA 5.1. *Let* $P \in \mathscr{A}$. *Then the optimal costs,* $C(B_m)$, $m = 1, 2, \cdots$, *converge to zero as* $m \to \infty$.

LEMMA 5.2. *Let* $P \in \mathscr{A}'$. *Then any sequence* $\bar{x}_m$, $m = 1, 2, \cdots$, *of trajectories admissible for* P_m, $m = 1, 2, \cdots$, *contains a subsequence that converges uniformly to a trajectory* x^*, *admissible for* P. *Moreover, all the limit points of any such sequence are admissible trajectories of* P.

Proofs of these lemmas are given in [22]. Lemma 5.1 is intuitively obvious for any consistent discretization. It is used in the proof of Lemma 5.2.

Proof of Theorem 5.1. Let $z_m = (\bar{x}_m, \bar{u}_m, \bar{v}_m, \bar{p}_m, \lambda_m, \beta_m)$, $m = 1, 2, \cdots$, be controls and multipliers obtained from Theorem 4.2 for the $\bar{x}_m$. (In this proof, unless specifically stated otherwise, the subscript m will be used to denote vector quantities corresponding to P_m.) Define μ_m as in Theorem 4.2.

For each m, by (4.8), $|p^0| + |p^1| + \cdots + |p^m| + |\beta| \neq 0$. But, in fact, if $p^m = 0$ for some m, then $w_j^m = 0, j = 1, \cdots, m$. Since $w_j^m = w_j^k, k = 0, 1, \cdots, m - 1$ and $j = 1, \cdots, m$, then $\mu_m(t) \equiv 0$ and hence, $y^j = 0$, $j = 0, 1, \cdots, m$. But, $y^0 = 0$ implies that $\beta = 0$. But, by (4.10), this is a contradiction of (A5). Hence, $p^m \neq 0$, $m = 1, 2, \cdots$. Since the multiplier conditions are homogeneous, without loss of generality set $|p^m| = 1, m = 1, 2, \cdots$. Hence,

$$|\mu_m(t)| \leqq \sum_{j=1}^{m} |w_j^m| \leqq 1, \qquad t \in I. \tag{5.1}$$

Therefore, since each μ_m is monotone nondecreasing and the functions μ_m, $m = 1, 2, \cdots$, are uniformly bounded, by Hellys' selection theorem [23] there exists a subsequence (w.l.o.g. denoted by m) and a bounded, nonpositive, monotone nondecreasing function μ^* such that $\mu_m(t) \to \mu^*(t)$, $t \in I$.

By definition, $\bar{p}_m = (\bar{y}_m, \bar{w}_m)$, where $\bar{y}_m$ satisfies equations (4.9). It is easy to prove that the piecewise linear extensions y_m of $\bar{y}_m$ are uniformly bounded, that $\dot{y}_m$ exists a.e. for each m, and that the $|\dot{y}_m|$ are uniformly bounded over $m = 1, 2, \cdots$. Therefore, the y_m are equicontinuous and by Arzela's theorem there exists a subsequence (w.l.o.g. denoted by m) and a function y^* of bounded variation on $I = [0, T]$ such that y_m converges uniformly to y^* on I.

Following Warga [24], given any $\varepsilon > 0$ there exists a set $I(\varepsilon)$ with measure greater than $T - \varepsilon$ such that μ_m, y_m, x_m converge uniformly to μ^*, y^*, x^*, respectively on $I(\varepsilon)$. Therefore, given any $h > 0$ there exists an $M(h)$ such that for $m > M(h)$ and $t \in I(\varepsilon)$, when $\dot{y}_m(t)$ exists,

$$\dot{y}_m(t) \in U(-f_x^T(x^*(t), U)y^*(t) - \mu^*(t)b_x(x^*(t), U), h). \tag{5.2}$$

Let $S(t, h)$ denote the set in (5.2), and let I' be a set of full measure on which $\dot{y}^*$ and $\dot{y}_m$, $m = 1, 2, \cdots$, exist. Each set $S(t, h)$ is convex and compact, each y_m is absolutely continuous, and there exists a K such that $|\dot{y}_m(t)| \leqq K$, $m = 1, 2, \cdots$, on I'. Therefore, for $t \in I'$,

$$\dot{y}^*(t) \in S(t, K\varepsilon). \tag{5.3}$$

But ε is arbitrary and the sets $S(t, K\varepsilon)$ are continuous functions of ε; hence,

$$\dot{y}^*(t) \in S(t) = -f_x^T(x^*(t), U)y^*(t) - \mu^*(t)b_x(x^*(t), U) \quad \text{a.e.} \tag{5.4}$$

(the set $g(x, U) = \{y | y = g(x, u), u \in U\}$). A similar argument demonstrates that

$$\dot{x}^*(t) \in f(x^*(t), U) \quad \text{a.e.}$$

Therefore, by Filippov's lemma there exists a control u^* admissible for P such that equations (4.3) are satisfied by x^*, u^*, μ^*, and z^*.

Proof of (4.2). First prove that $\mu^*(T)a(x^*(T)) = 0$. If $a(x^*(T)) = \gamma < 0$, then for large m, $v_m^m < \gamma/2$, where the subscript denotes the mth component of the mth component of $\bar{v}_m$. Hence, by (4.11) and (4.12), $\mu_m(T) = w_m^m = 0$ for large m.

Let $I_1 = (t_1, t_2) \subset I - Z$. Then there exists a sequence of closed intervals $I_n = [s_n, r_n]$ such that $s_n \downarrow t_1$ and $r_n \uparrow t_2$. Fix n and set $\gamma_n = \sup_{t \in I_n} a(x^*(t)) < 0$. There exist points of subdivision $t_m^{k(m)} \downarrow s_n$ and $t_m^{l(m)} \uparrow r_n$ as $m \to \infty$. (To simplify the notation the arguments of k and l are removed.) For large m, and $t \in J_m = [t_m^k, t_m^l]$, $v_j^m < \gamma_n/2, j = k, \cdots, l$ (j denotes the jth component of the mth component of $\bar{v}_m$). Hence, by (4.9), (4.11) and (4.12), $w_j^m = 0, j = k, \cdots, l$, and $\mu_m(t_m^k) = \mu_m(t_m^l)$. Hence, μ_m is constant on J_m for large m. But, $\mu_m(s_n) \to \mu^*(s_n)$, $\mu_m(r_n) \to \mu^*(r_n)$, and μ^* is monotone. Thus, μ^* is constant on I_n, for each n, and thus constant on I_1.

Proof of (4.4). Let $\gamma^*(t) = \langle z^*(t) + \mu^*(t)a_x(x^*(t)), f(x^*(t), u^*(t))\rangle$. (The dependence on t will be suppressed.) Define $\gamma_m = \langle \bar{y}_m + \bar{\mu}_m a(\bar{x}_m), f(\bar{x}_m, \bar{u}_m)\rangle$. (The quantities with bars are to be considered as the piecewise constant (left continuous) extensions of the corresponding discrete-valued functions to I.) In each case, these variables assume the value of the discrete variable at the left-hand endpoint of the subinterval in question, except $\bar{y}_m$ is the extension of the discrete-valued $\bar{y}_m$ to I that assumes the value of $\bar{y}_m$ at the right-hand endpoint. Let

$$q(z, \lambda, x) = \max_{u \in U} \langle z + \lambda a_x(x), f(x, u)\rangle.$$

Choose Q such that $q(\bar{y}_m, \bar{\mu}_m, \bar{x}_m) \leqq Q$ and $q(z^*, \mu^*, x^*) \leqq Q$ for all $t \in I$. Since y_m and x_m converge uniformly to z^* and x^* respectively; and by Egoroff's theorem, μ_m converges almost uniformly to μ^*, for any $\varepsilon > 0$ there exists a set $I(\varepsilon)$ with measure greater than or equal to $T - \varepsilon/Q$ such that for $t \in I(\varepsilon)$ and large m,

$$|q(z^*(t), \mu^*(t), x^*(t)) - q(\bar{y}_m(t), \bar{\mu}_m(t), \bar{x}_m(t))| \leqq \varepsilon/T. \tag{5.5}$$

From (4.13), the maximum condition

$$\gamma_m(t) = q(\bar{y}_m, \bar{\mu}_m, \bar{x}_m). \tag{5.6}$$

Hence, from (5.5) and (5.6), $\gamma_m(t) \to q(z^*(t), \mu^*(t), x^*(t))$ a.e. Since $\dot{x}_m$, $m = 1, 2, \cdots$, exists a.e. and is uniformly bounded over m, there exists a subsequence (w.l.o.g. denoted by m) and a function g such that $\dot{x}_m$ converges weakly to g. But x_m is absolutely continuous so (modulo $x_m(0)$ which is chosen to converge to $x^*(0)$),

$$x_m(t) = \int_0^t \dot{x}_m(s)\, ds \to \int_0^t g(s)\, ds = x^*(t)$$

and $g = \dot{x}^*$ a.e. Therefore, γ_m converges weakly to γ^*. Hence, $\gamma^* = q(z^*, \mu^*, x^*)$.

Proof of (4.1). Prove that $\gamma^*(t) \equiv |z^*(t)| + |\mu^*(t)| > 0$ for $t \in I$. Since z^* and μ^* appear in the linear equation (4.3), it is trivial to prove that there exists a K such that for any $t, \bar{t} \in I$,

$$|z^*(t)| \leqq K|z^*(\bar{t})| + K\left|\int_{\bar{t}}^t |\mu^*|\right|. \tag{5.7}$$

Let $\bar{t} = \inf\{t \in I | \gamma^*(t) = 0\}$. Suppose that $\bar{t} > 0$.

By (4.4) for a.e. $t \in I$ and all $u \in U$,

$$\langle z^* + \mu^* a_x(x^*), f(x^*, u^*)\rangle \geqq \langle z^* + \mu^* a_x(x^*), f(x^*, u)\rangle. \tag{5.8}$$

Therefore, for a.e. t and all $u \in U$,

$$\mu^* \frac{d}{dt}(a(x^*(t))) \geqq \mu^* b(x^*, u) + \langle z^*, f(x^*, u) - f(x^*, u^*)\rangle. \tag{5.9}$$

Hence, from (A4), (5.9), (5.7), and (4.2) there exists a K^1 such that for $t < \bar{t}$,

$$\frac{d}{dt} a(x^*(t)) \leqq \alpha + K^1|t - \bar{t}|.$$

Hence, for $t < \bar{t}$ and $|t - \bar{t}| < -\alpha/2K^1$,

$$a(x^*(\bar{t})) \leqq \frac{\alpha}{2}(\bar{t} - t) + a(x^*(t)) \leqq \frac{\alpha}{2}(\bar{t} - t) < 0.$$

But, μ^* is constant on any interval in $I - Z$. Hence, there exists $t^* < \bar{t}$ such that $\mu^*(t^*) = 0$. Since μ^k is nonpositive and monotone nondecreasing, $\mu^*(t) = 0, t^* \leqq T$. Hence, by (5.7), $z^*(t^*) = 0$, so $\gamma^*(t^*) = 0$, a contradiction. The only other possibility is $\bar{t} = 0$. Then $\mu^*(t) \equiv 0$. But, by construction, $1 = |y_m(T)| + |w_m^m| = |y_m(T)| + |\mu_m(0)|$ which converges to $|z^*(T)| + |\mu^*(0)| = 0$ as $m \to \infty$.

Proof of (4.5). First, consider G^0. By assumption [8, p. 86], the Jacobian $\partial g^0/\partial x$ has full rank, hence, $\beta_m \to \beta^*$ as $m \to \infty$. Therefore,

$$\bar{v}^*(0) \equiv \hat{z}^*(0) + \mu^*(0) a_x(x^*(0)) = -[\partial g^0/\partial x(x^*(0))]^T \beta_1^*,$$

where $\beta_1^* \in E^a$. Similarly, at G^1, from (4.11), $\lambda_m \to \lambda^*$ as $m \to \infty$ and

$$\hat{z}^*(T) = [\partial q^1/\partial x(x^*(T))]^T \lambda_1^*,$$

where $\lambda_1^* \in E^b$ and $\lambda_1^* \leqq 0$. Let $N = \{j | q_j^1(x^*(T)) = 0\}$. By (4.12), if $j \notin N$, the jth component of λ_1^* is zero. Therefore,

$$\langle \hat{z}^*(T), x - x^*(T) \rangle = \sum_{j \in N} \lambda_j^* \langle \partial q_j^1 / \partial x, x - x^*(T) \rangle. \tag{5.10}$$

But q_j^1 is convex and $q_j^1(x^*(T)) = 0$; hence each term in the summation is non-negative.

Proof of (4.6). If $a(x^*(T)) < 0$, then $\mu^*(T) = 0$ by (4.2); hence $v^*(T) = 0$ would imply $z^*(T) = 0$ contradicting (4.1).

If $a(x^*(T)) = 0$ and $v^*(T) = 0$, then

$$\hat{z}^*(T) = -\mu^*(T) a_x(x^*(T)). \tag{5.11}$$

By (5.10), for all $\hat{x} \in G^1$,

$$\langle \hat{z}^*(T), \hat{x} - \hat{x}^*(T) \rangle \geqq 0. \tag{5.12}$$

But a is convex, so for any $x \in A$,

$$\langle a_x(x^*(T)), \hat{x} - \hat{x}^*(T) \rangle \leqq a(x) \leqq 0,$$

contradicting (5.11) and (5.12) unless $x^*(T) \in \operatorname{int} G^1$, in which case $\hat{z}^*(T) = 0$. Thus $\mu^*(T) = 0$, contradicting (4.1). Therefore, $v^*(T) \neq 0$. Suppose there exists $t_n \uparrow T$ as $n \to \infty$ such that $v^*(t_n) = 0$. Clearly, this would imply that $\mu^*(t_n) \uparrow \gamma \leqq \mu^*(T)$ and

$$z^*(T) = -\gamma a_x(x^*(T)). \tag{5.13}$$

But (5.13) yields the same contradiction as (5.11). Therefore, there exists an interval $I_1 \subseteq I$ such that $v^*(t) \neq 0$ on I_1.

Proof of Corollary 5.1. The proof is immediate.

Proof of Corollary 5.2. The proof is trivial. The quantity $\tilde{\rho}_m$ appears only in (4.12).

Proof of Corollary 5.3. The cost functional of each problem D_m is not differentiable. This apparent difficulty can be circumvented by introducing the positive and negative components of ρ [8, p. 94]. Since $|\rho_m| \to 0$ as $m \to \infty$ and f is bounded, the effect of ρ on the necessary conditions of optimality vanishes as $m \to \infty$.

Proof of Corollary 5.4. Fix m. An application of [8, p. 91] to P_m results in a slight change in the relationships obtained in Theorem 4.2. In addition to $\bar{p}$, β, and λ there exists a $\gamma^0 \in E^c$ that appears in (4.9) when $k = 0$. That is,

$$y^0 - y^1 = A^0 y^1 \, \Delta t + \mu^0 b_x^0 \, \Delta t + [\partial q^0 / \partial x]^T \gamma^0. \tag{5.14}$$

As in Theorem 4.2, $|p^m| \neq 0$ for all m. Otherwise the independence of the gradients is contradicted; so the multipliers can be constructed as before. Moreover, for the same reason, since y_m^m and y_m^0 converge, the corresponding sequences λ_m, γ_m^0, and β_m also converge. In fact (see (4.9)), $\gamma_m^0 \to 0$ as $m \to \infty$, and the argument proceeds as before.

Remarks. A time optimal control problem can be converted into a fixed time problem, where the initial set, G^0, is defined by a set of equalities and inequalities. Hence, Corollary 5.4 states that Theorem 5.1 is applicable to such problems.

6. Sufficiency. Theorem 6.1 identifies a subclass $\mathscr{B}$ of $\mathscr{A}$ for which the conditions given in Theorem 4.1 are sufficient for optimality. If $P \in \mathscr{B}$, then by

Corollary 5.1, the optimal cost of P_m converges to the optimal cost of P as $m \to \infty$.

THEOREM 6.1. *Let $P \in \mathscr{A}$. Let x^* be an extremal of P generated from extremals of P_m for some infinite sequence m. If $G^1 = E^n$, and $a(x^*(T)) < 0$, and if*

$$\hat{f}(x, u, t) = C(t)\hat{x} + \hat{g}(u, t), \tag{6.1}$$

$$f_1(x, u, t) = h^1(\hat{x}, t) + g^1(u, t), \tag{6.2}$$

f_1 convex in $\hat{x}$ for each $(u, t) \in U \times I$,

$$G^0 = \{\hat{x}^0\}, \tag{6.3}$$

then x^* is an optimal solution of P.

Proof of Theorem 6.1. By (4.2), $\mu^*(T) = 0$. Since $G^1 = E^n$, $\hat{y}^m(T) = 0$ for $m = 1, 2, \cdots$. But, by (4.1), $|z^*(T)| + |\mu^*(T)| \neq 0$, hence, $z_1^*(T) < 0$. Let (x, u) denote any pair admissible for P. Since z^*, x, and x^* are absolutely continuous,

$$\langle z_1^*(T), (x_1(T) - x_1^*(T))\rangle = \langle \hat{z}^*(0), \hat{x}(0) - \hat{x}^*(0)\rangle - \langle \hat{z}^*(T), \hat{x}(T) - \hat{x}^*(T)\rangle + \int_0^T \frac{d}{dt}\langle z^*(t), x(t) - x^*(t)\rangle.$$

The first term is zero and from (4.5), the second term is nonnegative. Moreover,

$$\begin{aligned} d\langle z^*, x - x^*\rangle/dt &= \langle \dot{z}^*, x - x^*\rangle + \langle z^*, \dot{x} - \dot{x}^*\rangle \\ &= -\mu^*\langle b_x^*, x - x^*\rangle + z_1^* \cdot (h^1(x) - h^1(x^*) - \langle h_x^{1*}, x - x^*\rangle) \\ &\quad + \langle v^*, g(u) - g(u^*)\rangle - \mu^*\langle a_x, g(u) - g(u^*)\rangle. \end{aligned}$$

Since $z_1^* < 0$ and h^1 is a convex function of x, the second term is nonpositive. The maximum principle (4.4) states that the third term is nonpositive a.e. Moreover, the sum of the first and last terms equals $-\mu^* d\langle a_x^*, \hat{x} - \hat{x}^*\rangle/dt$. But μ^* is monotone and $\gamma(t) = \langle a_x^*, \hat{x} - \hat{x}^*\rangle$ is absolutely continuous; therefore,

$$-\int_0^T \mu^*\left(\frac{d\gamma}{dt}\right) dt = -\mu^*(T)\gamma(T) + \mu^*(0)\gamma(0) + \int_0^T \gamma(t)\, d\mu^*, \tag{6.4}$$

where the second integral is a Lebesgue–Stieltjes integral.

From (4.2), μ^* is constant on any interval $I_1 \subset I - Z$, and is monotone nondecreasing on I. Furthermore, $\gamma(t) \leqq 0$ for $t \in Z$. By definition the second integral in (6.4) equals

$$\lim_{n \to \infty} \sum_{k=0}^{n-1} \gamma(\xi_k^n)[\mu^*(t_{k+1}^n) - \mu^*(t_k^n)] \tag{6.5}$$

for any sequence of partitions of I such that $t_k^n \leqq \varepsilon_k^n \leqq t_{k+1}^n$, $k = 0, 1, \cdots, n$, $n = 1, 2, \cdots$, and the length of these subintervals converges to zero as $n \to \infty$.

For each n choose $s_k^n \in Z$ and $r_k^n \in I - Z$, $k = 1, \cdots, n$, such that $r_k \uparrow$ and $s_k \uparrow$ as $k \uparrow n$ and $|r_{k+1} - r_k| < 1/n$ and $|s_{k+1} - s_k| < 1/n$. Order these numbers by size, for example, $r_1 < r_2 < s_1 < r_3 < s_2 < s_3 < \cdots$ and consider the summation in (6.5). For each term there are three possibilities. If we have successive r_k, then since by (4.2), μ^* is constant on $[r_k, r_{k+1}]$, the corresponding term drops out of (6.5). If we have an r_k next to an s_j or two successive s_j, then define the corresponding

ξ to be s_j. Since $\mu^* \uparrow$ and $\gamma(\xi) \leqq 0$, the corresponding term in (6.5) is nonpositive. Therefore, the summation is nonpositive; hence the limit and thus the Stieltjes integral in (6.4) is nonpositive. To complete the argument consider the boundary term. If $T \in Z$, then $\langle a_x^*, \hat{x} - \hat{x}^* \rangle|_T \leqq a(\hat{x}) \leqq 0$, and if $T \notin Z$, then $\mu^*(T) = 0$; hence this term has the proper sign. To simplify the notation hats have been omitted frequently.

7. Heuristic discussion. Now reconsider the discrete problems $P_m^*, m = 1, 2, \cdots$, defined in § 3. For fixed m, minimize the first component of x^m over all $(\bar{x}_m, \bar{u}_m, \bar{v}_m)$ satisfying

$$x^{k+1} - x^k = f(x^k, u^k)\Delta t, \qquad k = 0, 1, \cdots, m-1, \tag{3.1}$$

$$\hat{x}^0 \in G^0, \quad u^k \in U, \qquad k = 0, 1, \cdots, m-1, \tag{3.2}$$

$$\hat{x}^m \in \cup(G^1, \rho_m), \quad a(x^k) \leqq \rho_m, \qquad k = 0, 1, \cdots, m, \tag{3.3}$$

where ρ_m is the minimum enlargement in the terminal and in the intermediate state constraints needed to obtain feasible solutions of (3.1) through (3.3). An application of [8, p. 91] yields the following theorem.

THEOREM 7.1. *Let $P \in \mathscr{A}$. Fix m. If $(\bar{x}_m, \bar{u}_m)$ is an optimal solution of P_m^*, then there exist costate vectors $p^0, p^1, \cdots, p^m$ in E^{n+1}, multipliers $\lambda^0, \lambda^1, \cdots, \lambda^m$ with $\lambda^k \in E^1$, $k = 0, 1, \cdots, m-1$, $\lambda^m \in E^{b+1}$, and $\mu \in E^a$ such that* (4.7), (4.8), *and* (4.10) *hold. Also*

$$p^k - p^{k+1} = A^k p^{k+1}\Delta t + a_x^k \lambda^k, \qquad k = 0, 1, \cdots, m-1, \tag{4.9'}$$

$$p^m = [\partial q^1/\partial x(x^m), a_x(x^m)]\lambda^m, \tag{4.11'}$$

$$\begin{aligned} \langle \lambda^k, a(x^k) \rangle &= 0, \qquad k = 0, 1, \cdots, m-1, \\ \langle \lambda^m, (q^1(x^m), a(x^m)) \rangle &= 0. \end{aligned} \tag{4.12'}$$

For $k = 0, 1, \cdots, m-1$,

$$\langle p^{k+1}, f(x^k, u^k) \rangle = \max_{u \in U} \langle p^{k+1}, f(x^k, u) \rangle. \tag{4.13'}$$

Observe that in (4.9') the gradient of the constraint function a_x appears in place of the gradient b_x, where $b = \langle a_x, f(x, u) \rangle$, that appears in (4.9). The coefficient of a_x in (4.9') varies with k in a totally disorganized fashion. Recall that the coefficient of b_x in (4.9) is a nondecreasing function of k and this fact is used to obtain a limiting multiplier μ^*. The use of limiting arguments in the present situation seems dubious. In fact, consider the function $v^* = z^* + \mu^* a_x(x^*)$ given in Theorem 4.1. Since μ^* is a bounded, monotone function, it is differentiable a.e. Hence, by (4.3),

$$\dot{v}^* = -f_x^T(x^*, u^*)v^* + \dot{\mu}^* a(x^*) \quad \text{a.e. } I. \tag{7.1}$$

Equation (7.1) indicates that in (4.9'), $\dot{\mu}^*$ is being approximated by $\bar{\lambda}_m/\Delta t$, and v^* is being approximated by $\bar{p}_m$. Clearly, $\dot{\mu}^*$ and v^* are not as well-behaved as μ^* and z^*. Hence, the relationship between P_m, $m = 1, 2, \cdots$, and P is stronger than that between P_m^*, $m = 1, 2, \cdots$, and P.

A situation similar to this occurs in Chang [19]. In [19] penalty functions were introduced to remove the constraint $x(t) \in A$, $t \in I$, thus generating a sequence of problems without intermediate state constraints. Chang applied the maximum principle to each of these problems and took limits of the multipliers. The theorem he obtained is essentially correct. Namely (under certain hypotheses), given any optimal pair (x^*, u^*) of P there exists a scalar function γ^* such that $\gamma^*(t) = 0$ if $x^*(t)$ is in the interior of A and $\gamma^*(t) > 0$ if $x^*(t)$ is on the boundary of A; and there exists a function $v^* : I \to E^{n+1}$ such that

$$\dot{v}^* = -f_x^T(x^*, u^*)v^* + \gamma^* a_x \quad \text{a.e.}, \tag{7.2}$$

$$\langle v^*, f(x^*, u^*)\rangle = \max_{u \in U} \langle v^*, f(x^*, u)\rangle \quad \text{a.e.} \tag{7.3}$$

Observe the connections between (7.2) and (4.9′), and between (7.3) and (4.13′). Chang is working with the complete multiplier $v^* = z^* + \mu^* a_x(x^*)$. This multiplier is not necessarily absolutely continuous and may not equal the integral of its derivative. Hence, difficulties in the limiting arguments exist and parallel those encountered in using P_m^*, $m = 1, 2, \cdots$, as an approximation to P.

8. A numerical example. As an application of a discrete approximation, consider the following example.

Example 2 (Speyer and Bryson [25]). Minimize

$$C(u) = \int_0^3 \frac{u^2(t)}{2}\, dt$$

subject to the constraints

$$\dot{h} = s, \quad h(0) = 1.6, \quad h(3) = 0,$$

$$\dot{s} = a, \quad s(0) = s(3) = -1,$$

$$\dot{a} = u, \quad a(0) = a(3) = 0,$$

and $s(t) \geqq -4t^2 + 12t - 8$ for $t \in [0, 3]$. Denote this problem by P^2.

Speyer and Bryson obtained a solution to this problem by using an extended version of Pontryagin's maximum principle. P^2 is of particular interest because the state constraint is a second order constraint [25]. The optimal trajectory consists of three sections, corresponding to the subintervals $[0, t_1)$, $[t_1, t_2)$, and $[t_3, 3]$. The first and third sections are interior arcs, and the second section is a boundary arc. However, the problem cannot be solved by simply solving for the unconstrained arcs separately and then piecing arcs together.

The construction presented in § 3 was used to solve P^2. Observe first that P^2 does not satisfy assumption (A4). Second, as stated earlier, it does not make sense to use the unconstrained formulations of B_m and P_m, $m = 1, 2, \cdots$. This formulation was introduced only for use in the proof of convergence. Since the desired relationships have been established, the formulation given in (3.1), (3.2), (3.4), (3.5) and (3.6) can be rearranged to any equivalent formulation. Moreover, since the system equations in P^2 are linear, the nonlinear terminal state constraint in (3.5) is particularly undesirable. However, Corollary 5.3 states that this constraint can be replaced by the introduction of an error into the difference equation at the last time interval, thus preserving the linearity.

Hence, equations (3.1) were used to obtain expressions for $\hat{x}_m^k = (h^k, s^k, a^k)$, $k = 1, \cdots, m$, in terms of the controls u^k, $k = 0, 1, \cdots, m-1$; and these expressions were then substituted into (3.4) to obtain similar expressions for each v^k ($v(t) \equiv -s(t) - 4t^2 + 12t - 8$). The resulting expressions for v^k were then substituted into the inequalities (3.6) to obtain equivalent inequalities involving the control u^k, $k = 0, 1, \cdots, m-1$. Thus, (3.1) and (3.4) were eliminated. The terminal constraints were handled by using the expressions for $\hat{x}^m = (h^m, s^m, a^m)$ in terms of the controls obtained from (3.1), and replacing the constraint $\hat{x}^m = \hat{x}^f$ by

$$|\hat{x}_i^m - \hat{x}_i^f| < \rho_i, \qquad i = 1, 2, 3.$$

Consequently, the inequalities used in the computations have the following form:

$$S^k(u) \leqq \rho_4, \qquad k = 1, \cdots, m, \tag{8.1}$$

$$|\hat{x}_i^m - \hat{x}_i^f| \equiv |R_i^k(u)| \leqq \rho_i, \qquad i = 1, 2, 3, \tag{8.2}$$

where $u = (u^0, \cdots, u^{m-1})$. Inequality (8.1) represents the intermediate state constraint on the speed, $s(t)$; and inequality (8.2) represents the terminal state constraints. In this particular example there are no control constraints. The position h and the acceleration a are constrained only by the difference equations. Satisfaction of the difference equations is guaranteed because the expressions for h^k, s^k, and a^k in terms of the controls that were used to reduce the other constraints to ones involving only the controls were obtained from these difference equations.

The trapezoidal rule was used to difference the differential equations instead of the simpler Cauchy–Euler explicit form used in the theoretical discussion. Obviously, the difference scheme chosen will affect the convergence of the iterates. The sample time $(\Delta t)_m$ was set equal to $3/m$, $m = 10 + 2q$, $q = 0, 1, \cdots, 10$.

For each m, the procedure consists of two steps.

Step 1 (*Feasibility*). Minimize

$$\sum_{i=1}^{4} \rho_i$$

over all u satisfying (8.1) and (8.2).

Step 2 (*Optimality*). Set each ρ_i in (8.1) and (8.2) equal to the value of the minimal sum obtained in Step 1 and minimize

$$\left[\sum_{j=0}^{m-2} \frac{u_j^2 + u_{j+1}^2}{2} + u_{m-1}^2\right]\frac{\Delta t}{2}$$

over all u satisfying the resulting inequalities (8.1) and (8.2).

For this particular example, Step 1 is a linear programming problem, essentially a phase-one procedure, with $m + 6$ rows and $3m + 10$ columns. Step 2 is a quadratic programming problem with $m + 6$ rows and $3m + 6$ columns.

In P^2, the cost functional is a norm on the control function u; hence, the optimal control functions (actually the piecewise constant or piecewise linear extensions of these functions to the interval $[0, 3]$) for the discrete problems $m = 1, 2, \cdots$ converge in the strong L_2-topology to the optimal control of P as $m \to \infty$. The corresponding state trajectories converge in the uniform topology to the optimal trajectory of P.

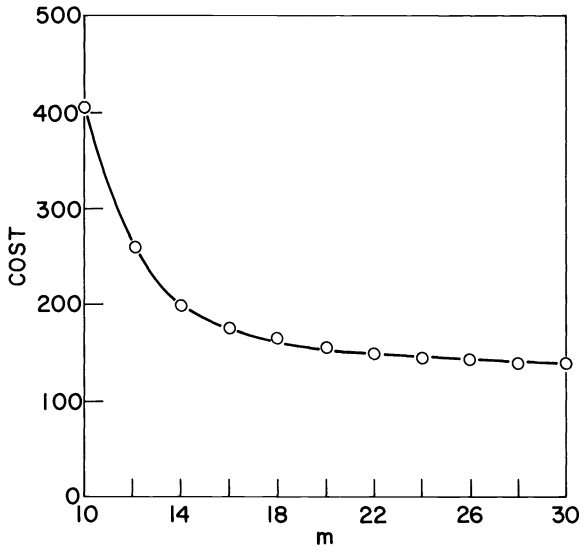


FIG. 1. *Cost vs m*

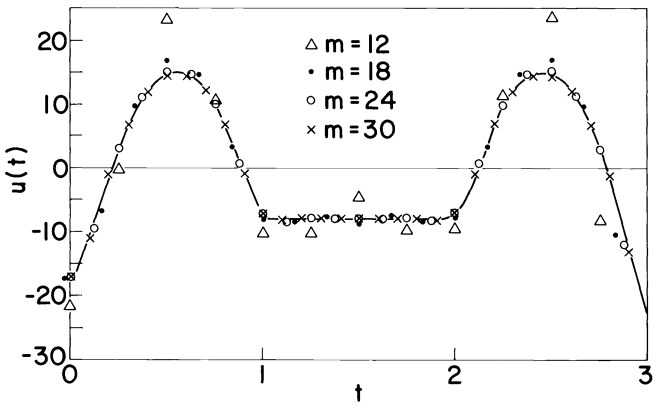


FIG 2. *Control vs time*

Numerical results are presented in Figs. 1, 2, and 3. Figure 1 depicts the convergence of the optimal cost of P_m as m increases. The behavior of the optimal control near the endpoints of the time interval affects the cost significantly. Figure 2 depicts the convergence of the optimal controls. Figure 3 depicts the convergence of the optimal acceleration. Figure 4 depicts Speyer and Bryson's results. From Figs. 2 and 3 one can see the difference in the convergence properties of the control and of the state variables. The results for the speed and the position, s and h, respectively, were not plotted because they converged more rapidly than the acceleration.

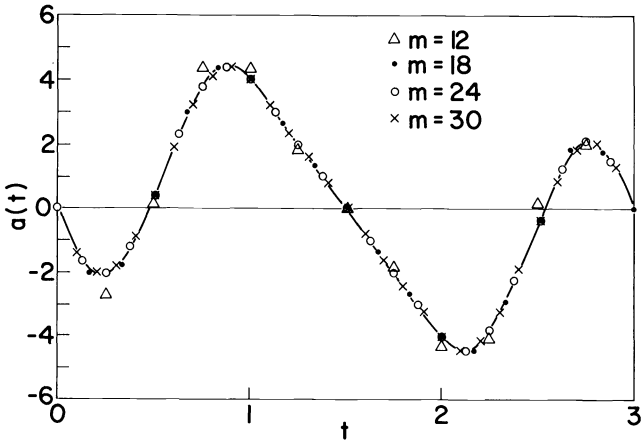


FIG. 3. *Acceleration vs time*

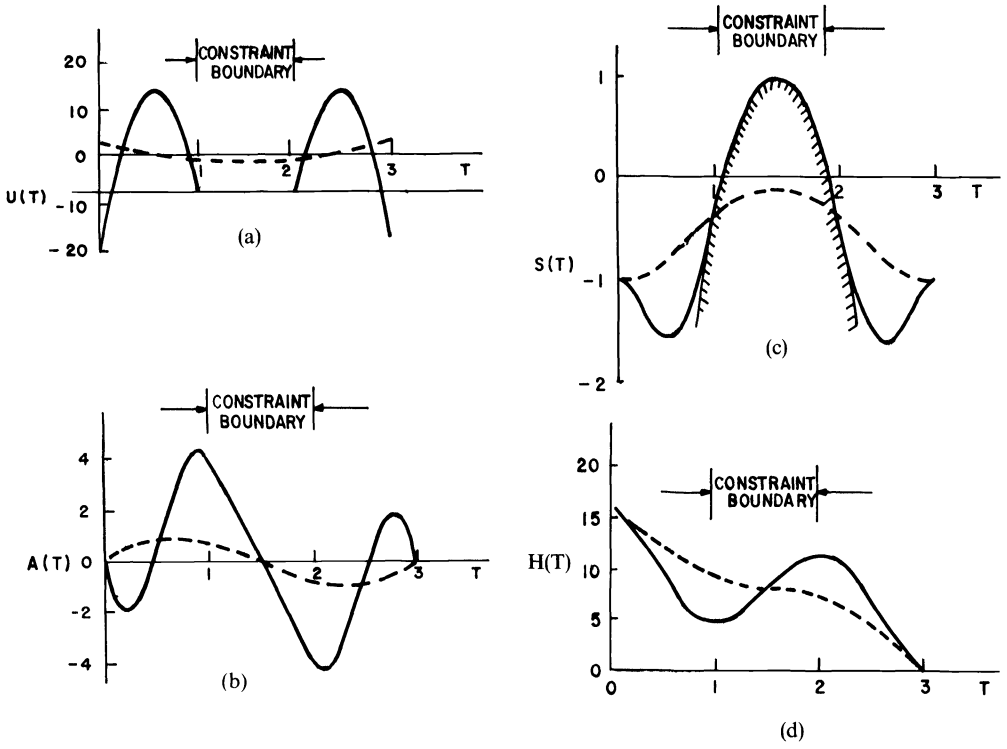


FIG. 4. *Speyer and Bryson results*: (a) *Control vs time*; (b) *Acceleration vs time*; (c) *Speed vs time*; (d) *Position vs time*

For $m = 12$ or $\Delta t = \frac{1}{4}$, the general shape of the optimal control was obtained but the peaks are too high. The values of the acceleration a for $m = 12$ are, however, fairly close to the actual optimal values. For $m = 18$ or $\Delta t = \frac{1}{6}$, the optimal control is fairly well identified, and all the values of the acceleration are very, very close to their optimal values.

REFERENCES

[1] J. B. ROSEN, *Optimal control and convex programming*, Proc. IBM Scientific Computing Symposium on Control Theory and its Applications (1964), IBM Data Processing Division, White Plains, N.Y., 1966, pp. 223–238.

[2] DANIEL TABAK, *Applications of mathematical programming techniques in optimal control; survey*, Proc. Joint Automatic Control Conference, Atlanta, Georgia, June 24–26, 1970, pp. 648–650.

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

[4] DANIEL TABAK, *A direct and nonlinear programming approach to the optimal nuclear reactor shutdown control*, IEEE Trans. on Nuclear Science, NS-15 (1968), pp. 57–59.

[5] R. R. MOHLER, *Optimal control computations for nuclear reactors*, Computing Methods in Optimization Problems, vol. 2, L. Zadeh, L. Neustadt and A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 257–270.

[6] R. MCGILL, *Optimal control, inequality state constraints, and the generalized Newton–Raphson algorithm*, this Journal, 3 (1965), pp. 291–298.

[7] J. L. SPEYER, R. K. MEHRA AND A. E. BRYSON, JR., *The separate computation of arcs for optimal flight paths with state variable inequality constraints*, Advanced Problems and Methods for Space Flight Optimization; Proc. Colloquium on Optimization in Space Flight Mechanics, University of Liege, Liege, Belgium, June 19–22, 1967.

[8] M. D. CANON, C. CULLUM AND E. POLAK, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.

[9] J. CULLUM, *Discrete approximations to continuous optimal control problems*, this Journal, 7 (1969), pp. 32–49.

[10] ———, *An explicit method for discretizing continuous optimal control problems*, J. Optimization Theory Appl., 8 (1971), pp. 15–34.

[11] L. NEUSTADT, *Discrete time optimal control systems*, International Symposium on Nonlinear Differential Equations and Nonlinear Mechanisms, Academic Press, New York, 1963, pp. 267–283.

[12] J. CULLUM, *Perturbations of optimal control problems*, this Journal, 4 (1966), pp. 473–487.

[13] B. M. BUDAK, E. M. BERKOVICH AND E. N. SOLOV'EVA, *Difference approximations in optimal control problems*, this Journal, 7 (1969), pp. 18–31.

[14] J. W. DANIEL, *On the convergence of a numerical method for optimal control problems*, J. Optimization Theory Appl., 4 (1969), pp. 330–342.

[15] J. WARGA, *Unilateral variational problems with several inequalities*, Michigan Math. J., 12 (1965), pp. 449–480.

[16] L. S. PONTRYAGIN, V. G. BOLTYANSKIĬ, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[17] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[18] L. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. II. Applications*, this Journal, 5 (1967), pp. 90–137.

[19] S. S. L. CHANG, *Optimal control in bounded phase space*, Automatica, 1 (1962), pp. 55–67.

[20] H. HALKIN, *Optimal control for systems described by difference equations*, Advances in Control Systems: Theory and Applications, C. Leondes, ed., Academic Press, New York, 1964, pp. 173–196.

[21] J. HOLTZMAN AND H. HALKIN, *Directional convexity and the maximum principle for discrete systems*, this Journal, 4 (1966), pp. 263–275.

[22] J. CULLUM, *Finite-dimensional approximations to state-constrained continuous optimal control problems*, IBM RC 3335, Yorktown Heights, N.Y., 1971.

[23] I. NATANSON, *Theory of Functions of a Real Variable*, vol. 1, Frederick Ungar, New York, 1964.

[24] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–145.

[25] J. SPEYER AND A. BRYSON, JR., *Optimal programming problems with a bounded state space*, AIAA J., 6 (1968), pp. 1488–1491.

[26] A. GOLDSTEIN, *On steepest descent*, this Journal, 3 (1965), pp. 147–151.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

A STABILITY THEORY FOR PERTURBED DIFFERENCE EQUATIONS*

S. P. GORDON†

Abstract. The problem of preserving stability properties under small perturbations for the solutions of difference equations is considered. The approach used is to study the behavior of the solutions of the perturbed difference equation with respect to the solutions of the original unperturbed difference equations. This leads to the introduction of notions which parallel the usual concepts of stability, asymptotic stability, instability and the like for the behavior of the perturbed solutions with respect to the unperturbed ones.

The principal technique employed is an extension of Lyapunov's direct method based on the difference of the two solutions. A series of theorems is obtained yielding criteria for each type of behavior for the perturbed solutions in terms of the existence of a discrete Lyapunov-type function with appropriate properties.

1. Introduction. The problem of preserving stability properties under small perturbations for the solutions of difference equations has never been adequately studied. Hahn [2] gave one result on exponential stability being preserved under a small perturbation of a linear system. The present author has developed a discrete analogue of total stability, along with several sufficient criteria for it in terms of the existence of discrete Lyapunov-type functions, in a separate paper [1].

In the present article, this problem is approached in a different manner. Instead of determining which stability properties are preserved under perturbations, the behavior of the solutions of the perturbed difference equation with respect to those of the original equation are studied. The results are obtained in terms of the existence of an extended form of the Lyapunov function.

2. Definitions and basic concepts. We shall consider the difference equation

$$X(n+1) = f(n, X(n)). \tag{1}$$

$f(n, X)$ here represents a function with values in E^m, an arbitrary m-dimensional vector space, and defined on some region D in $I \times E^m$ which contains the axis $\{X = 0, n \in I\}$, where I is the set of nonnegative integers. For simplicity, we may choose for D the semi-infinite cylinder

$$D_{n_0 R} = \{(n, X) \in I \times E^m : n \geqq n_0 \geqq 0, \|X\| \leqq R\}.$$

Here, $\|X\|$ denotes any m-dimensional norm of the vector X. In most cases, the upper bound R will be taken to be finite. The only exception would occur when we are considering the case of instability of the solutions of difference equations when the solutions become unbounded and hence the region under consideration must accommodate them.

The difference equation (1) will, in addition, be subject to the initial condition

$$X(n_0) = x_0.$$

* Received by the editors July 19, 1971.

† Department of Mathematics, Queens College of the City University of New York, Flushing, New York 11367.

The unique solution to the equation satisfying the given initial condition will be denoted by $F(n, n_0, x_0)$.

Finally, we note that if for a particular point (k, Z) in D_{n_0R}, for a finite value R,

$$\|f(k, Z)\| > R,$$

then obviously $Z(k + 1)$ is also larger in norm than R. Consequently, $Z(k + 2)$ is not defined by the equation (1). As a result of these remarks, unless otherwise stated, we shall concern ourselves solely in the sequel with those solutions to the difference equation which are defined for all $n \geqq n_0$ and which, in norm, do not exceed some finite value R. Equivalently, the only solutions considered are those which start in D_{n_0R} and remain in it for all $n \geqq n_0$. Moreover, we note that for these solutions, it is possible to write

$$\|f(n, X)\| \leqq R \quad \text{for all } n \geqq n_0.$$

In addition to the above difference equation (1), we also consider the associated perturbed difference equation

$$Y(n + 1) = f(n, Y(n)) + g(n, Y(n)), \tag{2}$$

where $g(n, Y)$ is also a function from D_{n_0R} into E^m. If the additional term $g(n, Y)$ is small in some sense, it is reasonable to expect that the behavior of the solutions of the perturbed equation will be similar to that for the solutions of the original equation, provided that the initial values for the two equations are sufficiently close.

The investigation will be carried out by using a certain class of real scalar functions $V(n, X)$ also defined on D_{n_0R} and satisfying the requirement

$$V(n, 0) = 0 \quad \text{for all } n \geqq n_0.$$

The following additional properties will also be required. Let M_0 represent the class of all real-valued monotone increasing functions $a(r)$ defined and positive for $r > 0$ and such that $a(0) = 0$. In terms of this, a real scalar function $V(n, X)$ is positive definite if there exists a function $a(r)$ of class M_0 such that

$$V(n, X) \geqq a(\|X\|) \quad \text{for all } n \geqq n_0.$$

A real scalar function $V(n, X)$ is positive semidefinite if $V(n, X) \geqq 0$ for all $n \geqq n_0$. Entirely similar definitions hold for such a function being either negative definite or negative semidefinite.

Furthermore, corresponding to a function $V(n, X)$, we define its total difference

$$\Delta V(n, X) = V(n + 1, f(n, X(n))) - V(n, X(n)).$$

$\Delta V(n, X)$ is obviously a measure of the growth or decay of the function $V(n, X)$ with regard to increasing n along the discrete trajectories represented by the solution of the difference equation (1). It should be noted that in general this can be calculated without direct knowledge of the actual solutions. Moreover, it will occasionally prove convenient, in a notational sense, to write

$$\Delta V(n, X) = V(n + 1, X(n + 1)) - V(n, X(n)).$$

We now introduce the types of possible behavior for the solutions of the perturbed difference equation (2) which will be of interest to us in the sequel.

DEFINITION 1. The solutions of the perturbed difference equation (2) are said to be *stable with respect to the unperturbed difference equation* (1) if, for all $\varepsilon > 0$, and for all $n_0 \in I$, there exists a $\delta(\varepsilon, n_0) > 0$ such that $\|y_0 - x_0\| < \delta$ implies that

$$\|G(n, n_0, y_0) - F(n, n_0, x_0)\| < \varepsilon$$

for all $n \geqq n_0$, for every solution $G(n, n_0, y_0)$ of the perturbed difference equation (2).

DEFINITION 2. The solutions of the perturbed difference equation (2) are said to be *quasi-asymptotically stable with respect to the unperturbed difference equation* (1) if, for all $n_0 \in I$, there exists a $\delta_0(n_0) > 0$ such that $\|y_0 - x_0\| < \delta_0$ implies that

$$\|G(n, n_0, y_0) - F(n, n_0, x_0)\| \to 0$$

as $n \to \infty$ for every solution $G(n, n_0, y_0)$ of the perturbed difference equation (2).

DEFINITION 3. The solutions of the perturbed difference equation (2) are said to be *asymptotically stable with respect to the unperturbed difference equation* (1) if both Definitions 1 and 2 hold.

The last definition is equivalent to the statement that all solutions of the perturbed difference equation which start sufficiently near to the initial value of the unperturbed solution remain close to it and eventually approach it. Furthermore, it is worth remarking that these notions are all uniform relative to the initial values x_0.

Moreover, we note that all of these definitions are independent of the behavior of the solutions of the unperturbed equation. In fact, the following simple examples show that the equilibrium of the original difference equation may be stable, asymptotically stable or even unstable.

Example 1. Consider the unperturbed difference equation

$$X(n + 1) = c,$$

where c is any constant, whose solution is $F(n, n_0, x_0) = c$, and the associated perturbed equation

$$Y(n + 1) = c + d,$$

for some sufficiently small constant d. The solution to this equation is given by

$$G(n, n_0, y_0) = y_0 = c + d.$$

As a consequence,

$$\|G(n, n_0, y_0) - F(n, n_0, x_0)\| = |d|$$

and Definition 1 holds with $\delta = \varepsilon$, for any $\varepsilon > 0$.

Example 2. Consider the unperturbed difference equation

$$X(n + 1) = aX(n),$$

where $|a| < 1$, whose asymptotically stable solution is given by

$$F(n, n_0, x_0) = a^{n - n_0} x_0.$$

Further, consider the perturbed difference equation

$$Y(n + 1) = (a + b)Y(n)$$

with b sufficiently small; in particular, take any b in the open interval $(0, 1 - a)$. The perturbed solution is then given by

$$G(n, n_0, y_0) = (a + b)^{n-n_0} y_0$$

and hence

$$\|G(n, n_0, y_0) - F(n, n_0, x_0)\| = \|(a + b)^{n-n_0} y_0 - a^{n-n_0} x_0\|,$$

which approaches zero as $n \to \infty$, for any y_0.

Example 3. Consider the unperturbed difference equation

$$X(n + 1) = X(n) + C,$$

where C is any constant, whose solution is given by

$$F(n, n_0, x_0) = x_0 + (n - n_0)C,$$

which is unstable. Thus, we would have the case in which R is unbounded. In addition, consider the perturbed equation

$$Y(n + 1) = Y(n) + C + g(n + 1),$$

where $\{g(n)\}$ is any sequence for which $\sum^{\infty} g(n) = 0$. The corresponding solution is then given by

$$G(n, n_0, y_0) = y_0 + (n - n_0)C + \sum_{k=n_0}^{n} g(k),$$

and, by the choice of $\{g(n)\}$, it is obvious that the difference between the two solutions approaches zero as n approaches infinity.

In addition, we also consider the following concept.

DEFINITION 4. The solutions of the perturbed difference equation (2) are said to be *unstable with respect to the unperturbed difference equation* (1) if, for every $\varepsilon > 0$ and every x_0 and every $n_0 \in I$, there exists some y_0 with $\|y_0 - x_0\| < \varepsilon$ such that

$$\|G(n_1, n_0, y_0) - F(n_1, n_0, x_0)\| \geqq \varepsilon \quad \text{for some } n_1 > n_0.$$

The above definition requires that for each solution to the unperturbed difference equation (1), a solution to the perturbed equation (2) can be found which starts arbitrarily close to the unperturbed solution and which eventually diverges from it.

3. Principal results. We now present several theorems which supply sufficient conditions for these types of behavior to hold in terms of the existence of real scalar functions $V(n, X)$.

THEOREM 1. *If there exists a real scalar function $V(n, X)$ such that, on $D_{n_0 R}$,*

(a) *$V(n, X)$ is continuous at $X = 0$,*

(b) *$V(n, X)$ is positive definite,*

(c) *$\Delta V(n, Y - X)$ is negative semidefinite;*

then the solutions of the perturbed difference equation (2) *are stable with respect to the unperturbed difference equation* (1), *provided that*

$$\|G(n, n_0, y_0) - F(n, n_0, x_0)\| \leqq R \quad \text{for all } n \geqq n_0.$$

Proof. Since $V(n, X)$ is positive definite, there is a function $a(r)$ of class M_0 such that

$$V(n, X) \geqq a(\|X\|).$$

Now, given any ε, choose y_0 sufficiently close to x_0 so that

$$\|y_0 - x_0\| < \varepsilon \quad \text{and} \quad V(n_0, y_0 - x_0) < a(\varepsilon).$$

It then follows that

$$\|G(n, n_0, y_0) - F(n, n_0, x_0)\| < \varepsilon \quad \text{for all } n \geqq n_0;$$

for, if not, there would be some $n_1 > n_0$ such that

$$\|G(n_1, n_0, y_0) - F(n_1, n_0, x_0)\| \geqq \varepsilon.$$

This, however, would imply that

$$\begin{aligned} V(n_1, G(n_1, n_0, y_0) - F(n_1, n_0, x_0)) &\geqq a(\|G(n_1, n_0, y_0) - F(n_1, n_0, x_0)\|) \\ &\geqq a(\varepsilon) \\ &> V(n_0, y_0 - x_0) \\ &\geqq V(n_1, G(n_1, n_0, y_0) - F(n_1, n_0, x_0)), \end{aligned}$$

which is a contradiction.

THEOREM 2. *If there exists a real scalar function $V(n, X)$ such that, on $D_{n_0 R}$,*

(a) *$V(n, X)$ is bounded below,*

(b) *$\Delta V(n, Y - X)$ is negative definite,*

then the solutions of the perturbed difference equation (2) *are asymptotically stable with respect to the unperturbed difference equation* (1), *provided that*

$$\|G(n, n_0, y_0) - F(n, n_0, x_0)\| \leqq R \quad \textit{for all } n \geqq n_0.$$

Proof. Since $\Delta V(n, Y - X)$ is negative definite, there exists a function $a(r)$ of class M_0 such that

$$\Delta V(n, Y - X) \leqq -a(\|Y - X\|).$$

Moreover, by a very straightforward inductive argument,

$$\begin{aligned} &V(n_0 + k, G(n_0 + k, n_0, y_0) - F(n_0 + k, n_0, x_0)) \\ &\quad = \sum_{j=0}^{k-1} [\Delta V(n_0 + j, G(n_0 + j, n_0, y_0) - F(n_0 + j, n_0, x_0))] + V(n_0, y_0 - x_0) \\ &\quad \leqq \sum_{j=0}^{k-1} [-a(\|G(n_0 + j, n_0, y_0) - F(n_0 + j, n_0, x_0)\|)] + V(n_0, y_0 - x_0). \end{aligned}$$

Taking the limit as $k \to \infty$ and using the fact that $V(n, X)$ is bounded below, we find that

$$\lim_{k \to \infty} \sum_{j=0}^{k-1} [a(\|G(n_0 + j, n_0, y_0) - F(n_0 + j, n_0, x_0)\|)] \leqq V(n_0, y_0 - x_0),$$

which implies that

$$a(\|G(n_0 + k, n_0, y_0) - F(n_0 + k, n_0, x_0)\|) \to 0 \quad \text{as } k \to \infty,$$

and therefore, since $a(r)$ is monotonically increasing,

$$\|G(n_0 + k, n_0, y_0) - F(n_0 + k, n_0, x_0)\| \to 0 \quad \text{as } k \to \infty.$$

It should be noted that both of these theorems hinge on the requirement

$$\|G(n, n_0, y_0) - F(n, n_0, x_0)\| \leqq R \quad \text{for all } n.$$

The following result gives one fairly simple set of conditions on the functions $f(n, X)$ and $g(n, Y)$ which will guarantee that this condition holds.

THEOREM 3. *If the function $f(n, X)$ satisfies a Lipschitz condition with respect to the variable X with constant $L < 1$, and if the function $g(n, X)$ satisfies*

$$\|g(n, X)\| \leqq a\|X\|$$

for some sufficiently small positive constant a, then

$$\|G(n, n_0, y_0) - F(n, n_0, x_0)\| \leqq R \quad \textit{for all } n \geqq n_0,$$

provided that y_0 is chosen sufficiently close to x_0.

Proof. For simplicity, we denote

$$f(n_0 + j) = f(n_0 + j, G(n_0 + j, n_0, y_0)).$$

It then follows, after a somewhat involved inductive argument, that

$$\begin{aligned}
&\|G(n_0 + k, n_0, y_0) - F(n_0 + k, n_0, x_0)\| \\
&\quad \leqq L^k\|y_0 - x_0\| + \|y_0\|[a^k + a^{k-1}L + a^{k-2}L^2 + \cdots + aL^{k-1}] \\
&\qquad + \|f(n_0)\|[a^{k-1} + a^{k-2}L + \cdots + aL^{k-2}] \\
&\qquad + \cdots + \|f(n_0 + k - 3)\|[a^2 + aL] + a\|f(n_0 + k - 2)\| \\
&\quad < L\|y_0 - x_0\| + R[(a^k + a^{k-1}L + \cdots + aL^{k-1}) \\
&\qquad + (a^{k-1} + a^{k-2}L + \cdots + aL^{k-2}) \\
&\qquad + \cdots + (a^2 + aL) + a] \\
&\quad = L\|y_0 - x_0\| + R[(a + a^2 + \cdots + a^k) + L(a + a^2 + \cdots + a^{k-1}) \\
&\qquad + \cdots + L^{k-1}(a)] \\
&\quad = L\|y_0 - x_0\| + Ra[(1 - a^k) + L(1 - a^{k-1}) + \cdots + L^{k-1}(1 - a)]/(1 - a) \\
&\quad < L\|y_0 - x_0\| + Ra[1 + L + \cdots + L^{k-1}]/(1 - a) \\
&\quad \leqq L\|y_0 - x_0\| + Ra/(1 - a)(1 - L).
\end{aligned}$$

This quantity, however, can be made smaller than R by taking y_0 sufficiently close to x_0 and by choosing the constant a sufficiently small, since $L < 1$.

By way of example, we now present one of the usual types of results on preserving stability under perturbations which is now merely an immediate application of Theorems 2 and 3.

THEOREM 4. *If the linear difference equation*

$$X(n + 1) = A(n)X(n)$$

is asymptotically stable with $\|A(n)\| \leqq b < 1$ for all n, then the solutions of the

perturbed difference equation

$$Y(n+1) = A(n)Y(n) + g(n, Y(n)),$$

where

$$\|g(n, Y)\| \leqq a\|Y\|$$

for some sufficiently small positive constant a, are also asymptotically stable.

THEOREM 5. *If there exists a real scalar function $V(n, X)$ such that, on D_{n_0R}:*

(a) *$V(n, X)$ is bounded;*

(b) *$\Delta V(n, Y - X) = aV(n, Y - X) + W(n, Y - X)$, where a is a positive constant and $W(n, X)$ is a semidefinite function defined on D_{n_0R};*

(c) *if $W(n, X)$ is not identically zero, then in each subdomain $D_{n_1r} \subset D_{n_0R}$ there exist points (n, X) for which $V(n, X)$ and $W(n, X)$ have the same sign for all $n \geqq n_1$;*

then the solutions of the perturbed difference equation (2) *are unstable with respect to the unperturbed difference equation* (1).

Proof. Suppose that D_{n_1r} is any subdomain of D_{n_0R} and that in it, $V(n, X)$ and $W(n, X)$ are positive at some points. Let (n_1, x_1) be one such point and choose any point (n_1, y_1) in D_{n_1r} at which $W(n, X)$ is positive. We consider the solutions $F(n, n_1, x_1)$ and $G(n, n_1, y_1)$ respectively of (1) and (2) for all $n \geqq n_1$. For convenience, we write them respectively as $F(n)$ and $G(n)$. Now consider

$$\begin{aligned}
&\Delta[(a+1)^{-n}V(n, G(n) - F(n))] \\
&\quad = (a+1)^{-(n+1)}V(n+1, G(n+1) - F(n+1)) - (a+1)^{-n}V(n, G(n) - F(n)) \\
&\quad = (a+1)^{-n}[\Delta V(n, G(n) - F(n))/(a+1) - aV(n, G(n) - F(n))/(a+1)] \\
&\quad = (a+1)^{-(n+1)}[\Delta V(n, G(n) - F(n)) - aV(n, G(n) - F(n))].
\end{aligned}$$

As a result,

$$\begin{aligned}
&\Delta V(n, G(n) - F(n)) - aV(n, G(n) - F(n)) \\
&\quad = (a+1)^{n+1}\Delta[(a+1)^{-n}V(n, G(n) - F(n))] \\
&\quad = W(n, G(n) - F(n)) \\
&\quad \geqq 0.
\end{aligned}$$

Consequently,

$$\Delta[(a+1)^{-n}V(n, G(n) - F(n))] \geqq 0$$

and hence the expression

$$(a+1)^{-n}V(n, G(n) - F(n))$$

increases with increasing n, for all $n \geqq n_1$. Thus,

$$(a+1)^{-n}V(n, G(n) - F(n)) \geqq (a+1)^{-n_1}V(n_1, G(n_1) - F(n_1)),$$

and so

$$V(n, G(n) - F(n)) \geqq (a+1)^{n-n_1}V(n_1, y_1 - x_1),$$

which becomes arbitrarily large as $n \to \infty$. However, by assumption, $V(n, X)$ is

bounded on $D_{n_0 R}$. Therefore, the sequence of points $\{(n, G(n) - F(n))\}$ must leave this domain as $n \to \infty$, and consequently,

$$\|G(n, n_1, y_1) - F(n, n_1, x_1)\| \geqq R$$

for all $n \geqq N$, for some N. Hence, the solutions of the perturbed difference equation (2) are unstable with respect to the unperturbed difference equation (1).

4. Concluding remarks. It is fairly apparent that the notions introduced in the present article can easily be extended to encompass as well the various refinements of the stability properties, such as uniform stability, equiasymptotic stability, uniform-asymptotic stability and so forth. Moreover, the author is currently in the process of developing the analogous theory for stability under perturbation for differential equations.

REFERENCES

[1] S. P. GORDON, *Stability and summability of solutions of difference equations*, Math. Systems Theory, 5 (1971), pp. 56–65.

[2] W. HAHN, *Über die Anwendung Methode von Liapunov auf Differenzengleichungen*, Math. Ann., 136 (1958), pp. 430–441.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

INVARIANCE AND STABILITY FOR BOUNDED UNCERTAIN SYSTEMS*

T. K. C. PENG†

Abstract. The positive limit sets of the solutions of a contingent differential equation are shown to possess an invariance property. In this connection the "invariance principle" in the theory of Lyapunov stability is extended to systems with unknown, bounded, time-varying parameters, and thus to a large and important class of nonautonomous systems. Asymptotic stability criteria are obtained and applied to guaranteed cost control problems.

1. Introduction. For ordinary differential equations an invariance principle has been developed by LaSalle [1], [2], [3] to obtain asymptotic stability results. The key idea is that for certain types of systems a positive limit set is an invariant set. This property was exploited for autonomous systems in [1] and for periodic systems in [2]. It was applied to "asymptotically autonomous" systems in [4] and the invariance property was extended to almost periodic systems in [5]. In this paper an invariance principle is established for the following class of uncertain systems:

$$\dot{x}(t) = f(x(t), q(t)), \quad x(0) = x_0, \tag{1}$$

where $x(t) \in R^n$, real n-space, $t \in R^1$. The function q is called *admissible* if it is measurable and if $q(t) \in \Omega$ for all t, where Ω is a compact set in R^n. The function f is continuous in (x, q), and is of class C^1 in x ($\partial f/\partial x$ is continuous in (x, q)). With an admissible q, the solution of (1) is defined to be the absolutely continuous function $\psi(t)$ which satisfies (1) almost everywhere on its interval of definition. It is also assumed that f is such that the solution of (1), for any admissible q and any x_0, has no finite escape time, at least on $[0, \infty)$.

Consider the set-valued function

$$R(x) \triangleq \{f(x, q) : q \in \Omega\}.$$

It is clear that $R(x)$ is compact for each x. Associated with (1), we can introduce the contingent equation

$$\dot{x}(t) \in R(x(t)), \quad x(0) = x_0. \tag{2}$$

A solution of (2) is defined to be an absolutely continuous function $\psi(t)$ such that $\psi(0) = x_0$ and $\dot{\psi}(t) \in R(\psi(t))$ almost everywhere on its interval of definition. It is known [6, p. 106] that ψ is a solution of (2) on a compact interval if and only if ψ is a solution of (1) for some admissible function q on the same interval. Thus a solution of (2) on $[0, \infty)$ or on $(-\infty, \infty)$ means a solution of (1), for some admissible q, on $[0, \infty)$ or on $(-\infty, \infty)$.

* Received by the editors June 8, 1971, and in revised form February 2, 1972.

† Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91103. This work was supported by the National Science Foundation under Grant GK-16017, and was completed while the author was with the Department of Electrical Sciences, State University of New York at Stony Brook.

DEFINITION. Let $\psi(t)$ be a solution of (2) on $[0, \infty)$. A point p in R^n is called a *positive limit point* of $\psi(t)$ if there is a time sequence $\{t_i\}$, such that $t_i \to \infty$ as $i \to \infty$, and $\psi(t_i) \to p$. The collection of positive limit points (of $\psi(t)$) is called a *positive limit set* (of $\psi(t)$).

It is well known that if the solution $\psi(t)$ is bounded on $[0, \infty)$, its positive limit set is nonempty, connected and compact. We shall show in § 2 that under certain conditions the limit set is an invariant set in a definite sense. The property leads to a Lyapunov theorem in § 3 for the asymptotic stability of system (1) or (2). In § 4 an example in [3] is reconsidered in a new perspective. In § 5 the theorem is applied to a guaranteed cost control problem.

2. Invariance property of limit sets. Let $\psi(t; q, x_0)$ denote the trajectory (or solution) of (1) starting at x_0 at time $t = 0$ with admissible function q.

DEFINITION. For system (1) or (2), a set N in R^n is called a *positively (negatively) invariant set* if, starting at any x_0 in N, there is an admissible function $q(t)$ on $[0, \infty)((-\infty, 0])$ such that the whole trajectory $\psi(t; q, x_0)$ stays in N for all $t > 0$ $(t < 0)$. The set N is called an *invariant set* if it is both positively invariant and negatively invariant.

DEFINITION. For system (1) or (2), a set N in R^n is called a *positively (negatively) quasi-invariant set* if, starting at any x_0 in N, there is a sequence of admissible functions $\{q^k(t)\}$ on $[0, \infty)$ $((-\infty, 0])$ such that, as $k \to \infty$, $\psi(t; q^k, x_0)$ converges to a function $\psi(t)$ uniformly on any interval $[0, T]$ $([-T, 0])$, while $\psi(t)$ remains in N for all $t > 0$ $(t < 0)$. The set N is called a *quasi-invariant set* if it is both positively quasi-invariant and negatively quasi-invariant.

LEMMA 1. *Assume that*:

(A1) *corresponding to each $T > 0$ and each bounded set B of R^n there is a bounded set $A(B, T)$ of R^n such that all solutions of* (2) *that start in B remain in $A(B, T)$ on $[0, T]$.*

Then if $\psi(t; q, x_0)$ is bounded as $t \to \infty$ for some admissible $q(t)$ on $[0, \infty)$, its positive limit set Γ^+ is a positively quasi-invariant set. If, in addition, it is assumed that:

(B) *the set $R(x)$ is convex for any x,*

then Γ^+ is a positively invariant set.

Proof. We shall first establish the existence of an absolutely continuous function $\bar{\psi}(t)$ on $[0, \infty)$ which stays in Γ^+, and then show that it is a limit curve for solutions of (2) under assumption (A1). Under the additional assumption (B), we shall show that $\bar{\psi}(t)$ is itself a solution of (2). Let $\bar{x} \in \Gamma^+$ and $T > 0$. There exist two sequences of functions $\{q^i(t)\}$, $\{\psi^i(t)\}$ defined on $[0, T]$ as follows:

$$q^i(t) \triangleq q(t_i + t), \tag{3}$$

$$\psi^i(t) \triangleq \psi(t; q^i, \bar{x}), \tag{4}$$

where $\{t_i\}$ is such that $t_i \to \infty$ and $\psi(t_i; q, x_0) \to \bar{x}$ as $i \to \infty$. By assumption (A1), there is a compact set $A(\bar{x}, T)$, such that all solutions of (2) starting at $\bar{x}$ stay in $A(\bar{x}, T)$. Therefore $\{\psi^i(t)\}$ are uniformly bounded on $[0, T]$. Also $\{\psi^i(t)\}$ are equicontinuous on $[0, T]$, because $\|f(x, q)\|$ (Euclidean norm) is continuous on the compact set $A(\bar{x}, T) \times \Omega$ and so $\|\dot{\psi}^i(t)\| \leqq \beta$ almost everywhere on $[0, T]$. By Ascoli's theorem, we can therefore assume that $\{\psi^i(t)\}$ converges uniformly to a function $\bar{\psi}(t)$ on $[0, T]$. Since $\|\dot{\psi}^i(t)\| \leqq \beta$, $\bar{\psi}(t)$ has a Lipschitz constant β

almost everywhere, and therefore is absolutely continuous on $[0, T]$. We next show that, for any $\tau \in [0, T]$, $\bar{\psi}(\tau)$ is a positive limit point of the original trajectory $\psi(t; q, x_0)$. Given any $\varepsilon > 0$ there exists $N_1 > 0$ such that for any $i > N_1$ and any $\tau \in [0, T]$,

$$\|\psi^i(\tau) - \bar{\psi}(\tau)\| < \varepsilon/2. \tag{5}$$

If we can show there exists $N_2 > 0$ such that for any $i > N_2$,

$$\|\psi(t_i + \tau; q, x_0) - \psi^i(\tau)\| < \varepsilon/2, \tag{6}$$

then it follows that for $i > \max\{N_1, N_2\}$,

$$\|\psi(t_i + \tau; q, x_0) - \bar{\psi}(\tau)\| < \varepsilon. \tag{7}$$

This means that $\bar{\psi}(\tau)$ is a positive limit point for $\psi(t; q, x_0)$. Therefore $\bar{\psi}(t) \in \Gamma^+$ for all $t \in [0, T]$.

To prove (6), simple continuity of solutions with respect to the initial conditions is not enough. We need to establish "equicontinuity" for solutions of (2) with respect to the initial conditions. Let $N(\bar{x}, d)$ be a neighborhood of $\bar{x}$ with radius d. By assumption (A1) there is a compact set $A(N, T)$ such that all solutions of (2) starting in $N(\bar{x}, d)$ stay in $A(N, t)$ for $t \in [0, T]$. Let $q^*(t)$ be admissible on $[0, T]$. Since $f(x, q^*(t))$ is C^1 in x, it is well known [8, p. 57] that $\psi(t; q^*, \xi)$ is C^1 in ξ for all $\xi \in N(\bar{x}, d)$ and, for almost all t on $[0, T]$, the following matrix equations hold:

$$\frac{\partial \psi}{\partial \xi}(0; q^*, \xi) = \text{identity matrix}, \tag{8}$$

$$\frac{d}{dt}\left[\frac{\partial \psi}{\partial \xi}(t; q^*, \xi)\right] = \frac{\partial f}{\partial x}(\psi(t; q^*, \xi), q^*)\frac{\partial \psi}{\partial \xi}(t; q^*, \xi), \tag{9}$$

where

$$\left[\frac{\partial \psi}{\partial \xi}\right]_{ij} \triangleq \frac{\partial \psi_i}{\partial \xi_j}, \qquad \left[\frac{\partial f}{\partial x}\right]_{ij} \triangleq \frac{\partial f_i}{\partial x_j}, \qquad i, j = 1, \cdots, n.$$

For a matrix F, let $\|F\| \triangleq \sup_{\|x\|=1} \|Fx\|$. Since $\|(\partial f/\partial x)(x, q)\|$ is continuous on $A(N, T) \times \Omega$, there is $\gamma > 0$ such that

$$\left\|\frac{\partial f}{\partial x}(x, q)\right\| \leqq \gamma \quad \text{for all } q \in \Omega, \quad x \in A(N, T). \tag{10}$$

From (9), (10) and the definition of $A(N, T)$, it follows that

$$\left\|\frac{\partial \psi}{\partial \xi}(t; q^*, \xi)\right\| \leqq e^{\gamma T} \tag{11}$$

for any $\xi \in N(\bar{x}, d)$, any admissible q^* and any $t \in [0, T]$. Let $\xi^1 \in N(\bar{x}, d)$. By the mean value theorem for functions on a vector space [9, p. 117], there exist η^i on the line segment joining ξ^1 and $\bar{x}$, such that

$$\psi_i(t; q^*, \xi^1) - \psi_i(t; q^*, \bar{x}) = \left[\frac{\partial \psi_i}{\partial \xi}(t; q^*, \xi)\right]_{\xi=\eta^i}(\xi^1 - \bar{x}), \qquad i = 1, \cdots, n, \tag{12}$$

where

$$\frac{\partial \psi_i}{\partial \xi} \triangleq \left[\frac{\partial \psi_i}{\partial \xi_1}, \cdots, \frac{\partial \psi_i}{\partial \xi_n}\right],$$

a row vector. Since $\eta^i \in N(\bar{x}, d)$, from (11), (12) and the fact that $\|\partial \psi_i/\partial \xi\| \leqq \|\partial \psi/\partial \xi\|$, we have

$$|\psi_i(t; q^*, \xi^1) - \psi_i(t; q^*, \bar{x})| \leqq e^{\gamma T}\|\xi^1 - \bar{x}\|, \qquad i = 1, \cdots, n. \tag{13}$$

From (13) it follows that

$$\|\psi(t; q^*, \xi^1) - \psi(t; q^*, \bar{x})\| \leqq \sqrt{n}\, e^{\gamma T}\|\xi^1 - \bar{x}\|. \tag{14}$$

The inequality (14) is true for any $\xi^1 \in N(\bar{x}, d)$, any $t \in [0, T]$ and any admissible q^*.

Let

$$0 < \delta < \min\left\{d, \frac{1}{2\sqrt{n}} e^{-\gamma T}\right\}.$$

We then have, for any $\xi \in N(\bar{x}, \delta)$ and any $t \in [0, T]$,

$$\|\psi(t; q^*, \xi) - \psi(t; q^*, \bar{x})\| < \varepsilon/2 \tag{15}$$

for any admissible q^*, especially for those q^i defined in (3). Since $\psi(t_i; q, x_0) \to \bar{x}$ as $i \to \infty$, there is $N_2 > 0$ such that for any $i > N_2$,

$$\|\psi(t_i; q, x_0) - \bar{x}\| < \delta. \tag{16}$$

From (3) and the uniqueness of the solution we have

$$\psi(t_i + \tau; q, x_0) = \psi(\tau; q^i, \psi(t_i; q, x_0)). \tag{16a}$$

Making use of (16) in (15) for ξ and setting $q^* = q^i$, we obtain (6) through (16a) and (4). It is thus proved that $\bar{\psi}(t) \in \Gamma^+$ for all $t \in [0, T]$. If assumption (B) is satisfied, it is easily verified by using Filippov's proof for the existence of an optimal control ([7] or [6, pp. 105–108]), that $\bar{\psi}(t)$ is a solution of (2) for some admissible $\bar{q}(t)$ on $[0, T]$. Starting at $\bar{\psi}(T)$ at time T, the definitions of $\bar{\psi}(t)$ and $\bar{q}(t)$ can be extended to the interval $[T, 2T]$ in exactly the same way as in $[0, T]$. By induction $\bar{\psi}(t)$ and $\bar{q}(t)$ are defined on $[0, \infty)$ with $\bar{\psi}(t) \in \Gamma^+$ and $\bar{q}(t)$ admissible. Therefore Γ^+ is a positively invariant set for (2). If assumption (B) is not satisfied, $\bar{\psi}(t)$ is nevertheless a solution of the equation

$$\dot{x} \in H(R(x)) \quad \text{on } [0, \infty), \tag{2a}$$

where $H(R(x))$ is the convex hull of the set $R(x)$. It is a "sliding state" of Gamkrelidze and, on any finite interval, is the uniform limit of solutions of (2) (see [13] or [8]). For Γ^+ to be positively quasi-invariant we now construct an approximating sequence for the sliding state $\bar{\psi}(t)$ on the semi-axis $[0, \infty)$. The following notation will be needed. Let $\psi[t; q, x_1, t_1]$ denote the unique solution of (1), for some admissible q, satisfying $\psi[t_1; q, x_1, t_1] = x_1$. Note that if $t_1 = 0$, this notation has been more conveniently written as $\psi(t; q, x_1)$. We again make use of uniqueness to obtain

$$\begin{aligned} \psi[t; q, x_1, t_1] &= \psi[t - t_1; q_{t_1}, x_1, 0] \\ &= \psi(t - t_1; q_{t_1}; x_1), \end{aligned} \tag{17a}$$

where

$$q_{t_1}(t) \triangleq q(t + t_1). \tag{17b}$$

For each $m \geqq 1$, let $\{q_m^i(t)\}$ be a sequence of admissible functions on $[(m-1)T, mT]$ such that as $i \to \infty$,

$$\|\psi[t; q_m^i, \bar{\psi}(mT - T), (m-1)T] - \bar{\psi}(t)\| \to 0 \tag{17c}$$

uniformly on $[(m-1)T, mT]$. For each j, define $q^j(t)$ on $[0, \infty)$ such that

$$q^j(t) \triangleq q_m^j(t) \quad \text{on } [(m-1)T, mT), \qquad m = 1, 2, 3, \cdots. \tag{17d}$$

Then $\{q^j(t)\}$ is a sequence of admissible functions on $[0, \infty)$. From (17c) and (17d), $\|\psi(t; q^j, \bar{x}) - \bar{\psi}(t)\| \to 0$ uniformly on $[0, T]$. In other words, given any $\varepsilon > 0$ there is $N_3 > 0$ such that for any $j > N_3$ and any $t \in [0, T]$,

$$\|\psi(t; q^j, \bar{x}) - \bar{\psi}(t)\| < \varepsilon. \tag{17e}$$

For $t \in [T, 2T]$, it follows from the uniqueness of the solution and (17d), (17a) that

$$\begin{aligned} \psi(t; q^j, \bar{x}) &= \psi[t; q^j, \psi(T; q^j, \bar{x}), T] \\ &= \psi(t - T; (q_2^j)_T, \psi(T; q^j, \bar{x})). \end{aligned} \tag{17f}$$

From (15) and the fact that $\psi(T; q^j, \bar{x}) \to \bar{\psi}(T)$ as $j \to \infty$, there exists $N_4 > 0$ such that for $j > N_4$, $t \in [T, 2T]$ and any admissible q^*,

$$\|\psi(t - T; q^*, \psi(T; q^j, \bar{x})) - \psi(t - T; q^*, \bar{\psi}(T))\| < \varepsilon/2. \tag{17g}$$

In particular, (17g) is satisfied for $q^* = (q_2^j)_T$. It therefore follows from (17f) that for $j > N_4$ and $t \in [T, 2T]$,

$$\|\psi(t; q^j, \bar{x}) - \psi[t; q_2^j, \bar{\psi}(T), T]\| < \varepsilon/2. \tag{17h}$$

Let $m = 2$ in (17c). There exists $N_5 > 0$ such that for $j > N_5$ and $t \in [T, 2T]$,

$$\|\psi[t; q_2^j, \bar{\psi}(T), T] - \bar{\psi}(t)\| < \varepsilon/2. \tag{17i}$$

From (17h) and (17i), (17e) is satisfied for $t \in [T, 2T]$ if $j > \max\{N_4, N_5\}$. Thus if $j > \max\{N_3, N_4, N_5\}$, then (17e) is satisfied for any $t \in [0, 2T]$. Repeating this process, (17e) is seen to hold on $[0, mT]$ for any $m \geqq 1$ as $j \to \infty$. This completes the proof.

A stronger result on the set Γ^+ is obtained if an assumption on system (2) stronger than (A1) is satisfied.

LEMMA 2. *Assume that*:

(A2): *Corresponding to each $T > 0$ and each bounded set B of R^n there is a bounded set $A(B, T)$ of R^n such that all solutions of* (2), *in both directions, that start in B at $t = 0$ remain in $A(B, T)$ on $[-T, T]$.*

Then, if $\psi(t; q, x_0)$ is bounded as $t \to \infty$ for some admissible $q(t)$ on $[0, \infty)$, its positive limit set Γ^+ is a quasi-invariant set. If assumption (B) *is also satisfied, then Γ^+ is an invariant set.*

Proof. Since assumption (A2) implies (A1), Γ^+ is positively quasi-invariant and, with assumption (B), Γ^+ is positively invariant. With (A2) the two important

properties of the solutions of (2), boundedness on finite intervals and equi-continuity with respect to initial conditions, hold true in both directions. We thus can show that Γ^+ is negatively quasi-invariant. Since the set $-R(x) \triangleq \{y: -y \in R(x)\}$ is convex if $R(x)$ is convex, we can also show that Γ^+ is negatively invariant, if assumption (B) is also satisfied. In fact, the proof follows the proof of Lemma 1 step by step, with all the time intervals replaced by their images on the negative real axis (for example, $[0, T]$ replaced by $[-T, 0]$), and with (3) defined for i sufficiently large.

Remark. A sufficient condition for assumption (A1) is

$$x'f(x, q) \leqq \alpha_1(1 + \|x\|^2) \tag{18a}$$

for all $q \in \Omega$ and $x \in R^n$. The existence of a constant α_1 in (18a) implies that

$$\|x(t)\|^2 \leqq (1 + \|x_0\|^2)\, e^{2\alpha_1 T} \quad \text{for all } t \in [0, T]. \tag{18b}$$

A sufficient condition for assumption (A2) is

$$|x'f(x, q)| \leqq \alpha_1(1 + \|x\|^2) \tag{18c}$$

for all $q \in \Omega$ and $x \in R^n$. In this case, (18b) is satisfied for all $t \in [-T, T]$.

3. Asymptotic behavior.

DEFINITION. Let $V(x)$ be a real C^1-function on R^n, and let G be any set in R^n. $V(x)$ is called a *Lyapunov function* for (2) on G if there is a nonnegative continuous function $W(x)$, such that

$$\frac{\partial V(x)}{\partial x} f(x, q) \leqq -W(x) \leqq 0 \quad \text{for all } q \in \Omega,\ x \in G. \tag{19}$$

We observe that (19) means that $(d/dt)V(x(t)) \leqq -W(x(t))$ almost everywhere along any trajectory.

Remark. For system (2), the existence of $W(x)$ is guaranteed by $(\partial V(x)/\partial x) f(x, q) \leqq 0$, for all $x \in G$, $q \in \Omega$, because the function

$$W_1(x) \triangleq -\max_{q \in \Omega} \left[\frac{\partial V(x)}{\partial x} f(x, q) \right] \geqq 0$$

is continuous. Any continuous positive lower bound of $W_1(x)$ is also qualified. One will appear in the guaranteed cost control problem in § 5.

Let $E = \{x: W(x) = 0, x \in \bar{G}\}$, where $\bar{G}$ is the closure of G. Let M be the largest invariant set and M_1 be the largest quasi-invariant set of (2) contained in E. Also, let M^+ be the largest positively invariant set and M_1^+ be the largest positively quasi-invariant set of (2) contained in E. It is clear that $M \subset M^+$ and $M_1 \subset M_1^+$. For simplicity we shall only deal with sets M and M_1 under assumption (A2) in the following theorems and corollaries. But whenever (A2) is not satisfied but (A1) can be assumed, we can simply replace M with M^+ and M_1 with M_1^+, and the resulting statements will still hold true.

Let $d(x, F) \triangleq \inf_{y \in F} (\|x - y\|)$ be the distance between a point x and a set F in R^n. We shall understand "$x(t) \to F$" to mean "$d(x(t), F) \to 0$."

THEOREM 1. *Under assumption* (A2), *if* $V(x)$ *is a Lyapunov function for* (2) *on* G, *then each bounded solution of* (2) *which remains in* G *as* $T \to \infty$ *converges to* M_1

as $t \to \infty$. *If assumption* (B) *is also satisfied, then each bounded solution converges to* M.

Proof. Let $x(t)$ be such a solution. Since it is bounded, its positive limit set Γ^+ is nonempty and $x(t) \to \Gamma^+$ as $t \to \infty$. Because of (19) it is proved in Theorem 1 of [10] that $\Gamma^+ \subset E$. By Lemma 2, Γ^+ is itself a quasi-invariant set. Therefore $\Gamma^+ \subset M_1$, which further implies $x(t) \to M_1$ as $t \to \infty$. If assumption (B) is satisfied, then Γ^+ is also invariant. Then $\Gamma^+ \subset M$ and $x(t) \to M$ as $T \to \infty$.

COROLLARY 1.1. *If* $V(x)$ *is a Lyapunov function on* R^n $(G = R^n)$, *then every bounded solution converges to* M_1 *as* $t \to \infty$. *If the set* $R(x)$ *is convex, then every bounded solution converges to* M.

We can apply many boundedness criteria (see [11]) to obtain sharper results. The following corollary is an example.

COROLLARY 1.2. *Under assumption* (A2), *if* $V(x)$ *is a Lyapunov function on* $R^n(G = R)$, *and if* $V(x) \geqq a(\|x\|)$, *where* $a(r)$ *is a continuously increasing function and* $a(r) \to \infty$ *as* $r \to \infty$, *then every solution of* (2) *converges to* M_1 *as* $t \to \infty$. *If the set* $R(x)$ *is convex, then every solution converges to* M.

Proof. By Theorem 10.1 of Yoshizawa [11, p. 38], every solution of (2) is bounded. Corollary 1.1 completes the proof.

Asymptotic stability of a state requires that it be both *stable* and an *attractor*. A set Q in R^n is said to be an attractor for solutions of (2) if there is a neighborhood N of Q such that for any q, $x_0 \in N$, this implies $\psi(t; q, x_0) \to Q$ as $t \to \infty$. We then have an immediate consequence of Theorem 1.

COROLLARY 1.3.[1] *Assume that* (A2) *and* (B) *are satisfied. Let* G *be a bounded, open, "absolutely" positive invariant set (all solutions of* (2) *starting in* G *remain in* G *for all* $t \geqq 0$). *If* V *is a Lyapunov function for* (2) *on* G *and* $M \subset G$, *then* M *is an attractor and* G *is in its region of attraction.*

Proof. By Theorem 1, each solution starting in G approaches M. Now M is the largest invariant set in E and is therefore closed (by the equicontinuity of solutions of (2) with respect to the initial conditions described by (15), it is easy to prove that the closure of an invariant set of (2) is invariant). Hence G is a neighborhood of M, and M is an attractor.

In Corollary 1.3 if $M = \{0\}$ and if it is known that $V(x)$ is positive definite ($V(x) > 0$ for $x \neq 0$, $V(0) = 0$), then $\{0\}$ is stable, and one has asymptotic stability. The region of attraction may vary with q but always contains G, so that G is an "absolute" estimate of the "extent" of the stability.

4. An example: Stability of nonautonomous systems.

Example 1.

$$(*) \qquad \dot{x}_1 = x_2, \quad \dot{x}_2 = -x_1 - p(t)x_2, \quad \text{where } p(t) \geqq \delta > 0.$$

This example was considered in [3] and the following argument was given. Define $V = \frac{1}{2}(x_1^2 + x_2^2)$. Then $\dot{V} = -p(t)x_2^2 \leqq -\delta x_2^2$. V is therefore a Lyapunov function on R^2. The best result is that any solution converges to the axis $x_2 = 0$; and since $V \to c$ as $t \to \infty$, $x_1 \to c_0$ as $t \to \infty$ (c_0 depends on p). If $p(t)$ is also bounded from above, $0 < \delta \leqq p(t) \leqq m$, then since $x_2(t) \to 0$ the system becomes

[1] This is closely parallel to Corollary 1 of [10].

"asymptotically autonomous" to the system

$$(**) \qquad \dot{x}_1 = x_2, \quad \dot{x}_2 = -x_1.$$

Solutions of (*) converge (by a theorem of Yoshizawa in [4]) to the largest invariant set of (**) (in the sense of autonomous systems, see [1]) contained in the set $x_2 = 0$, which is the origin. By Theorem 1 of this paper, the example is seen in a different light: the system with $0 < \delta \leqq p(t) \leqq m$ satisfies assumptions (A2) and (B) of Lemma 2; since the largest invariant set (in the present sense) of the original system (*) contained in the set $x_2 = 0$ is the origin, every solution converges to the origin. This argument is simpler than the previous one. It is direct and bypasses the "asymptotically autonomous" requirement, and therefore does not have to refer to system (**) which is difficult to obtain in many cases.

5. Application to guaranteed cost control problem.

5.1. Guaranteed cost control problem. The problem has been discussed in [12]. A brief description of the subject related to stability is the following. Given a control system with uncertain parameters:

$$\dot{x} = f(x, \eta(x), q), \quad x(0) = x_0, \tag{20a}$$

where $\eta(x)$ is a feedback control, an m-vector-valued function; q is an uncertain parameter, a measurable function of time with $q(t) \in \Omega$, a compact subset of R^r (therefore an admissible function as defined in (1)). Assume (20a) has a unique solution with no finite escape time on $[0, \infty)$ for any admissible q so that a cost functional can be defined:

$$C[x_0, \eta, q] \triangleq \lim_{T \to \infty} \int_0^T g(x(t), \eta(x(t)), q(t))\, dt, \tag{20b}$$

where g is a nonnegative scalar function. The functional C represents the cost associated with the trajectory characterized by x_0, η and q.

DEFINITION. A nonnegative scalar function $V(x_0)$ is called a *guaranteed cost* at x_0 for (20a), (20b), if the choice of $\eta(x)$ is such that

$$C[x_0, \eta, q] \leqq V(x_0) \tag{21}$$

for all admissible q. In this case $\eta(x)$ is called a *guaranteed cost control.*

A sufficient condition for $V(x)$ to be a guaranteed cost function at any x is the following lemma.

LEMMA 3. *If the row vector $\partial V(x)/\partial x$ of a scalar nonnegative C^1-function $V(x)$ satisfies*

$$\frac{\partial V(x)}{\partial x} f(x, \eta(x), \omega) + g(x, \eta(x), \omega) \leqq 0 \tag{22}$$

for all $x \in R^n$, $\omega \in \Omega$, then $V(x)$ is a guaranteed cost at any x, with guaranteed cost control η.

Proof. Let x_0 be an arbitrary initial condition, let q be an arbitrary admissible function and let $\psi(t)$ be the solution of (20a) with x_0, q and η. Integrate (22) along

the trajectory $\psi(t)$ from 0 to T. We have

$$0 \leqq \int_0^T g(\psi(t), \eta(\psi(t)), q(t))\, dt \leqq V(x_0) - V(\psi(T)) \leqq V(x_0). \tag{23}$$

Since T is arbitrary, $V(x_0)$ is a guaranteed cost at x_0.

5.2. Asymptotic stability. To discuss stability, let us adopt the following notation:

$$f_\eta(x, q) \triangleq f(x, \eta(x), q), \tag{24a}$$

$$g_\eta(x, q) \triangleq g(x, \eta(x), q). \tag{24b}$$

We can also define the set $R_\eta(x)$ for (24a) and the contingent equation $\dot{x} \in R_\eta(x)$ for the equation $\dot{x} = f_\eta(x, q)$, as in (2).

The idea of the following theorem is essentially the same as a theorem in [12] which was stated without proof.

THEOREM 2. *Let $f_\eta(x, q)$ be C^1 in x, continuous in q, and satisfy assumption* (A2) (*and* (B)). *Let $g_\eta(x, q)$ be continuous in x and q (g_η is nonnegative by* (24b)). *If $V(x)$ is a guaranteed cost function defined everywhere on R^n which is C^1 and satisfies inequality* (22), *then $V(x)$ is a Lyapunov function of the closed loop uncertain system* (20a) *on R^n, and every bounded solution, whatever admissible q, converges to the set $M_1(M)$ as $t \to \infty$, where $M_1(M)$ is the largest quasi-invariant (invariant) set for equation $\dot{x} \in R_\eta(x)$ contained in the set $E = \{x : W_2(x) = 0, x \in R^n\}$, and*

$$W_2(x) \triangleq \min_{q \in \Omega} [g_\eta(x, q)] \geqq 0. \tag{25}$$

Proof. Since g_η is continuous in both x and q, $W_2(x)$ is continuous. From (22) and (25),

$$\frac{\partial V(x)}{\partial x} f_\eta(x, q) \leqq -g_\eta(x, q) \leqq -W_2(x) \quad \text{for all } x \in R^n, q \in \Omega.$$

Hence $V(x)$ is a Lyapunov function for $\dot{x} \in R_\eta(x)$ on R^n. By Theorem 1 and Corollary 1.1, we immediately have Theorem 2.

5.3. Linear regulator problem with quadratic performance. The simplest guaranteed cost control problem is the following:

$$f(x, \eta(x), q) = (A(q) + BK)x, \tag{26a}$$

$$g(x, \eta(x), q) = \tfrac{1}{2}x'(H'H + K'QK)x, \tag{26b}$$

where $\eta(x) = Kx$, $Q > 0$ (positive definite), and q is the only time-dependent data, an uncertain admissible function. Matrix A is assumed to be continuous in q. In this case we can seek the solution of (22) in the form $V(x) = \frac{1}{2}x'Rx$, where R is taken to be symmetric and positive semidefinite. Then (22) becomes

$$R(A(\omega) + BK) + (A(\omega) + BK)'R + H'H + K'QK \leqq 0 \tag{27}$$

for all $\omega \in \Omega$. From Theorem 2 we have the following corollary.

COROLLARY 2.1. *If R satisfies* (27) *and $R > 0$, then every trajectory of the closed loop system* (26a), *whatever admissible q, converges to M_1, the largest quasi-invariant set contained in E, where*

$$E = \{x : x'(H'H + K'QK)x = 0, x \in R^n\}. \tag{28}$$

If the set $R_\eta(x) = \{A(q)x : q \in \Omega\}$ is convex, then the trajectory also converges to M, the largest invariant set in E.

Proof. Since $V(x) = \frac{1}{2}x'Rx \geqq \frac{1}{2}\lambda \min(R)\|x\|^2$, where $\lambda \min(R)$ is the smallest eigenvalue of R (therefore positive), the assertion follows from Theorem 2 and Corollary 1.2.

The set M shrinks to the origin under observability.

LEMMA 4. *For the linear system* (26a), (26b), $M = \{0\}$, *the origin, if and only if the linear system $S(q)$, defined by*

$$\dot{x} = (A(q(t)) + BK)x, \tag{29a}$$

$$y = \bar{H}x, \tag{29b}$$

is completely observable at $t = 0$ for any admissible $q(t)$, where

$$\bar{H}'\bar{H} \triangleq H'H + K'QK. \tag{30}$$

Proof. It is obvious that the origin belongs to M. Let $\Phi_q(t, \tau)$ be the state transition matrix of (29a) for some admissible q. It is well known that $S(q)$ is completely observable at $t = 0$ if and only if there is a $t > 0$ such that

$$U[q, t, 0] \triangleq \int_0^t \Phi_q'(\tau, 0)\bar{H}'\bar{H}\Phi_q(\tau, 0)\, d\tau > 0. \tag{31}$$

Sufficiency. Assume $S(q)$ is completely observable at $t = 0$ for any q, but $x_0 \neq 0$ belongs to M. Since $x_0 \in M$, there is an admissible q^1 such that the trajectory lies in E and $\bar{H}\Phi_{q^1}(t, 0)\, x_0 = 0$ for all $t \geqq 0$. But since $U[q^1, t_1, 0] > 0$ for some t_1, $\bar{H}\Phi_{q^1}(t, 0)x_0 \neq 0$ for some $t \in [0, t_1]$. This is a contradiction.

Necessity. Assume that $M = \{0\}$ but $S(q^*)$ is not completely observable at $t = 0$. Then $\det\{U[q^*, t, 0]\} = 0$ for all $t \geqq 0$. Let

$$N_t \triangleq \{x : x \in R^n, \|x\| = 1, x'[q^*, t, 0]x = 0\}.$$

Then N_t is a compact set and nonempty for all t; also $N_{t_1} \supset N_{t_2}$ for $t_2 > t_1$. Define $N \triangleq \bigcap_{0 \leqq t} N_t$. Then N is nonempty. Let $\xi \in N$. Then for all $t \geqq 0$,

$$\xi' U(q^*, t, 0)\xi = 0 \tag{32}$$

implying $\bar{H}\Phi_{q^*}(t, 0)\xi = 0$ for all $t \geqq 0$, which means $\xi \in M$. Since $\|\xi\| = 1$, it contradicts the assumption $M = \{0\}$.

A useful observation is the following lemma.

LEMMA 5. *A sufficient condition for $S(q)$ to be completely observable for all admissible q at $t = 0$ is*

$$\operatorname{rank}\begin{bmatrix} \bar{H} \\ \bar{H}(A(\omega) + BK) \end{bmatrix} = n \quad \text{for all } \omega \in \Omega. \tag{33}$$

Proof. Assume that $S(q^*)$ is not completely observable at $t = 0$ for admissible q^*; then as in the necessity proof of Lemma 4, there is ξ with unit norm such that

$\bar{H}\Phi_{q^*}(t, 0)\xi = 0$ for all $t \geqq 0$. Take the time derivative: $\bar{H}(A(q^*(t)) + BK)\Phi_{q^*}(t, 0)\xi = 0$ almost everywhere on $[0, \infty)$. By (33), since $\Phi_{q^*}(t, 0)$ is nonsingular, $\xi = 0$, a contradiction.

As an illustration we have the following example.

Example 2. For the linear system (26a), (26b), let $A(q(t)) = A_0 + q(t)A_1$, where

$$A_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad |q(t)| \leqq 1,$$

and $H = (1, 0, 0)$, $B = (0, 0, 1)'$, $Q = 1$. To find a solution R for (27) with a suitable K, we choose $K = -Q^{-1}B'R$ and compute the symmetric, positive semidefinite solution of the Riccati-type equation:

$$RA_0 + A_0'R + H'H - RBQ^{-1}B'R + N(R) = 0, \tag{34}$$

where $N(R) \triangleq S|\Lambda|S'$, $S\Lambda S' \triangleq RA_1 + A_1'R$ (S and Λ are orthogonal and diagonal matrices respectively), and $|\Lambda|$ is a matrix taking absolute values of the elements in Λ. Since $N(R) \geqq q(RA_1 + A_1'R)$ for all $|q| \leqq 1$, it is easily seen that (34) implies (27).

The numerical computation yields:

$$R = \begin{bmatrix} 2.543 & 3.236 & 1.124 \\ 3.236 & 7.112 & 2.861 \\ 1.124 & 2.861 & 3.643 \end{bmatrix} > 0, \quad K = (-1.124, -2.861, -3.643).$$

We shall show that (33) is satisfied. Observe that in this case, $\bar{H}'\bar{H} \triangleq H'H + K'K$ is of rank 2. Therefore, $\bar{H}y = 0$ if and only if $y = \alpha\xi$, where α is a scalar and $\xi = (0, 3.643, -2.861)'$. Assume that (33) is violated for some $\omega_1 \in [-1, 1]$. Then there is a $y \in R^n$, $y \neq 0$, such that $\bar{H}y = 0$ and $\bar{H}(A(\omega_1) + BK)y = 0$. This implies that $y = \alpha\xi$ for some $\alpha \neq 0$, and $(A(\omega_1) + BK)y = \beta\xi$ for some scalar β. The first component of $\beta\xi$ should equal zero. But $[(A(\omega_1) + BK)y]_1 = 3.643\alpha$ and $\alpha \neq 0$, a contradiction. By Corollary 2.1, since convexity holds, all solutions converge to M. By Lemmas 4 and 5, $M = \{0\}$. Therefore, the origin is asymptotically stable in the large for the closed loop system with any uncertain admissible function q.

Acknowledgment. The author is grateful to Professor S. S. L. Chang for his helpful suggestions, and to the referee for his helpful comments on an earlier manuscript.

REFERENCES

[1] J. P. LaSalle, *Some extensions of Liapunov's second method*, IRE Trans. Circuit Theory, CT-7 (1960), pp. 520–527.

[2] ———, *Asymptotic stability criteria*, Proc. Symposia Applied Mathematics, vol. 13, Hydrodynamic Instability, Amer. Math. Soc., Providence, 1962, pp. 299–307.

[3] ———, *An invariance principle in the theory of stability*, Proc. International Symposium (Puerto Rico), Academic Press, New York, 1967, pp. 277–286.

[4] T. YOSHIZAWA, *Asymptotic behavior of solutions of a system of differential equations*, Contributions to Differential Equations, 1 (1963), pp. 371–387.
[5] R. MILLER, *On almost periodic differential equations*, Bull. Amer. Math. Soc., 70 (1964), pp. 792–795.
[6] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
[7] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.
[8] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[9] T. M. APOSTOL, *Mathematical Analysis*, Addison-Wesley, Reading, Mass., 1957.
[10] J. P. LASALLE, *Stability theory for ordinary differential equations*, J. Differential Equations, 4 (1968), pp. 57–65.
[11] T. YOSHIZAWA, *Stability Theory by Liapunov's Second Method*, Math. Soc. Japan, Tokyo, 1966.
[12] S. S. L. CHANG AND T. K. C. PENG, *Adaptive guaranteed cost control of systems with uncertain parameters*, IEEE Trans. Automatic Control, AC-17 (1972), no. 4.
[13] R. V. GAMKRELIDZE, *Optimal sliding states*, Soviet Math. Dokl., 3 (1962), pp. 559–562.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

STATE CONSTRAINTS IN CONVEX CONTROL PROBLEMS OF BOLZA*

R. TYRRELL ROCKAFELLAR†

Abstract. Methods of convex analysis are applied to certain problems of Lagrange and Bolza in optimal control. Conditions characterizing optimal arcs are obtained without the usual differentiability assumptions on the data in the problem. Special existence theorems are proved. A dual control problem is formulated in terms of adjoint arcs which are not necessarily absolutely continuous, but of bounded variation, so as to allow for jumps caused by the presence of state contraints in the primal problem.

1. Introduction. Convex problems of Bolza are problems of a basic type in the calculus of variations and optimal control which satisfy convexity conditions not only in the control or derivative variables, but also in the state variables. Such problems have many special properties not implied by standard theory, particularly as regards duality. Moreover, these properties can be deduced by convexity methods which do not require the customary assumptions of continuity and differentiability. A number of results in this vein have been obtained by the author in [10], [11], [12]. However, these results do not explicitly treat problems with bounded state variables, and indeed they often exclude state constraints, other than constraints on endpoints. The purpose of this paper is to show how the results can nevertheless be applied to many problems with bounded state variables by various devices. The effect of state constraints on duality, on the existence of solutions, and on necessary and sufficient conditions for optimality is analyzed in detail.

We take as our model a "convex" control problem of the form:

$$\text{(1.1)} \qquad \text{Minimize} \quad \int_0^1 f(t, x(t), u(t))\,dt + l(x(0), x(1))$$

subject to

$$\text{(1.2)} \qquad \dot{x}(t) = A(t)x(t) + u(t) \quad \text{for almost every } t,$$

$$\text{(1.3)} \qquad x(t) \in X(t) \quad \text{for every } t,$$

where $x:[0,1] \to R^n$ is absolutely continuous and $u:[0,1] \to R^n$ is measurable. The sets $X(t)$ in R^n are nonempty, closed and convex, while the functions $f(t,\cdot,\cdot)$ and l on $R^n \times R^n$ are lower semicontinuous, convex and extended-real-valued—they may take on $+\infty$, although not $-\infty$, as a possible value, but they are assumed not to be identically $+\infty$. Of course, $A(t)$ denotes an $n \times n$ real matrix. The dependence of $X(t)$, $f(t,\cdot,\cdot)$ and $A(t)$ on t is discussed below. No differentiability is assumed.

It should be emphasized that, despite appearances, our problem of Bolza includes as special cases many other types of problems, such as problems of Lagrange

* Received by the editors October 13, 1971.

† Department of Mathematics, University of Washington, Seattle, Washington 98105. This research was supported in part by the Air Force Office of Scientific Research under Grant AF-AFOSR-71-1994.

with fixed endpoints. In fact, other constraints, besides the abstract constraint (1.3), are implicitly incorporated into the problem through the use of $+\infty$. Thus there is the endpoint constraint

$$(x(0), x(1)) \in E, \tag{1.4}$$

where E is the convex subset of $R^n \times R^n$ which is the effective domain of l:

$$E = \{(c_0, c_1) | l(c_0, c_1) < +\infty\}. \tag{1.5}$$

Similarly, there is the implicit control constraint

$$u(t) \in U(t, x(t)) \quad \text{for almost every } t, \tag{1.6}$$

where

$$U(t, x) = \{u | f(t, x, u) < +\infty\} \subset R^m. \tag{1.7}$$

The reader unfamiliar with this method of representing constraints is referred to [10] for further discussion and examples.

The implicit control set $U(t, x)$ is convex, but possibly of less than full dimension in R^m; thus the absence of a matrix $B(t)$ in (1.2) does not mean a loss of generality. (Note that $U(t, x)$ is not necessarily bounded or even closed.) The functions $f(t, x, \cdot)$ or l might in particular vanish identically on the sets $U(t, x)$ or E, respectively. The latter sets and the sets $X(t)$ might be described by inequality constraints. However, such specific situations need not concern us in the development of the basic theory. They can be handled at a later stage by a routine application of standard theorems in convex analysis. This is discussed in [10] in the case of $U(t, x)$ and E, and the considerations are similar for $X(t)$. Thus we can concentrate on the main features and difficulties of the problem, relegating many distracting and notationally burdensome details to "computation" in particular examples.

If $U(t, x)$ were empty for certain values of t and x, such values would have to be avoided, and this could amount to an implicit state constraint in addition to (1.3). We eliminate this possibility through our assumptions in the next section, our aim here being to keep the state constraints explicit and separate from other aspects of the problem, so that their exact role can be seen. However, there are problems which cannot be treated adequately in this way, because there is no sense to $U(t, x)$ being nonempty for $x \notin X(t)$. For example, in mathematical economics one encounters the case where $X(t)$ is the nonnegative orthant of R^n, and there is no way to define $f(t, x, u)$ finitely outside of this orthant without violating our convexity assumptions. Such problems require a different approach, involving a generalization of the basic duality theory to problems where the trajectories are not absolutely continuous, but only of bounded variation.

The plan of the paper is as follows. In § 2 and § 3 we derive lower semicontinuity properties of the cost functional and a basic existence theorem. These results may be compared most closely perhaps with those of Cesari [1] and Olech [9]; however, there is not a large overlap, and certainly the methods used here are very different. A dual control problem, involving the costate variables, is introduced in § 4 and studied in relation to certain optimality conditions in § 5. The optimality conditions, in a slightly generalized form, are shown in § 6 to be necessary and

sufficient under fairly mild hypotheses. The costate functions in these generalized conditions are characterized in § 7 as solutions to a generalized dual problem.

For comparison with other literature on necessary and sufficient conditions in problems with state constraints, papers of Neustadt [6], [7], [8] and Funk and Gilbert [3] may be cited. (Further references may be found in these papers.) For the most part, these authors focus on results applicable also to "nonconvex" problems, but which involve stronger regularity assumptions than ours (differentiability, etc.), even when specialized to the "convex" case. Neustadt's paper [8] does bring convexity to the fore, in methodology as well as hypothesis, and is thus the closest in spirit to the present work. However, also in [8] differentiability assumptions intervene, and the formulation of the optimality conditions is dependent on them. The conditions that we formulate do not even require the differentiability of the functions defining the state constraints, although, of course, this property could be exploited in analyzing the conditions in special cases.

No one has previously shown that the costate functions in the optimality conditions for problems with state constraints solve a general dual problem. The case of problems without state constraints was covered in our earlier papers [10], [12]. Some related results were also obtained under stronger regularity assumptions by Tsvetanov [16]. For an economic example not entirely covered by the results in this paper, for reasons mentioned above, see Makarov [17].

2. The Bolza functional. Let $\mathscr{A}$ denote the Banach space consisting of all absolutely continuous functions $x:[0, 1] \to R^n$ under the norm

$$\|x\|_{\mathscr{A}} = |x(0)| + \int_0^1 |\dot{x}(t)|\, dt, \tag{2.1}$$

and let $\mathscr{C}$ denote the larger Banach space consisting of all continuous functions $x:[0, 1] \to R^n$ under the usual norm

$$\|x\|_{\mathscr{C}} = \max_{0 \leqq t \leqq 1} |x(t)|. \tag{2.2}$$

(Here $|\cdot|$ denotes the Euclidean norm in R^n.) We have $\|x\|_{\mathscr{C}} \leqq \|x\|_{\mathscr{A}}$ for all $x \in \mathscr{A}$.

Let

$$S = \{x \in \mathscr{A} | x(t) \in X(t) \quad \text{for every } t\}. \tag{2.3}$$

Clearly S is a closed convex subject of $\mathscr{A}$. Our problem can be represented as:

$$\text{Minimize} \quad F(x) \quad \text{subject to} \quad x \in S, \tag{2.4}$$

where

$$F(x) = \int_0^1 f(t, x(t), \dot{x}(t) - A(t)x(t))\, dt + l(x(0), x(1)). \tag{2.5}$$

We call F a *Bolza functional* on the space $\mathscr{A}$. Of course, conditions must be imposed so that the integral in (2.5) makes sense.

INTEGRABILITY ASSUMPTION. *The components of the matrix $A(t)$ are summable as functions of $t \in [0, 1]$. Furthermore, $h(t, x, p)$ is (finite and) summable as a function*

of $t \in [0, 1]$ for each fixed $x \in R^n$ and $p \in R^n$, where

$$h(t, x, p) = \sup \{u \cdot p - f(t, x, u) | u \in U(t, x)\}. \tag{2.6}$$

The function h is not only convex in p, but also concave in x, by virtue of the joint convexity of $f(t, x, u)$ in x and u [13, Thm. 33.1].

The constraint $u \in U(t, x)$ could be omitted from (2.6) without loss of generality, in view of (1.7). The condition that $h(t, x, p) > -\infty$ for all (t, x, p) amounts to the condition, mentioned earlier, that $U(t, x) \neq \varnothing$ for all (t, x). The rest of the assumption on $h(t, x, p)$ is a growth condition of Nagumo–Tonelli type on $f(t, x, u)$ as a function of $u \in U(t, x)$. This is essential in obtaining the existence of solutions to the control problem, as is seen in the next section. It is also convenient technically in a great many other respects.

THEOREM 1. *Under the integrability assumption, F is a well-defined, lower semicontinuous, convex functional from the Banach space $\mathscr{A}$ to $R^1 \cup \{+\infty\}$. Moreover, the (convex) level sets*

$$\{x \in \mathscr{A} | F(x) \leqq \alpha\}, \quad \alpha \text{ real}, \tag{2.7}$$

are locally compact relative to the weak topology on $\mathscr{A}$. These sets are also closed and locally compact as subsets of the Banach space $\mathscr{C}$, with respect to both the weak and the strong topologies on $\mathscr{C}$.

Proof. This follows mainly from results in [12]. In the notation of [12], we have

$$F(x) = \Phi_{l,L}(x), \tag{2.8}$$

where

$$L(t, x, v) = f(t, x, v - A(t)x). \tag{2.9}$$

Condition (A) of [12] is satisfied by l and L, in view of our convexity and lower semicontinuity assumptions on l and $f(t, \cdot, \cdot)$. The Hamiltonian function corresponding to L is

$$\begin{aligned} H(t, x, p) &= \sup \{v \cdot p - L(t, x, v) | v \in R^n\} \\ &= h(t, x, p) + p \cdot A(t)x. \end{aligned} \tag{2.10}$$

By duality, we also have [10, p. 211]

$$L(t, x, v) = \sup \{v \cdot p - H(t, x, p) | p \in R^n\}. \tag{2.11}$$

The integrability assumptions on h and A imply that $H(t, x, p)$ is finite and summable in $t \in [0, 1]$ for each fixed $x \in R^n$ and $p \in R^n$. Hence L also satisfies conditions (B), (C_0) and (D_0) of [12] by the corollary and remark after Proposition 4 of [12]. These conditions guarantee in particular that F is well-defined, convex and lower semicontinuous [10, Thm. 1]. The local weak compactness of the level sets of F in $\mathscr{A}$ is a consequence of (C_0), as noted in [12, discussion following Thm. 1]. The last assertion of the theorem is obtained from the following fact.

LEMMA 1. *If a convex set K in $\mathscr{A}$ is closed and weakly locally compact, then it is also closed and locally compact as a subset of $\mathscr{C}$, both weakly and strongly.*

Proof. We observe first that if a set K_0 is weakly compact in $\mathscr{A}$, then it is strongly compact as a subset of $\mathscr{C}$. In fact, the set of function $\dot{x}$, as x ranges over K_0,

is weakly compact in the L^1-space of R^n-valued functions on $[0, 1]$. Hence, by the Dunford–Pettis criterion for weak compactness in L^1-spaces, there exists for every $\varepsilon > 0$ some $\delta > 0$ such that

$$\int_T |\dot{x}(t)|\, dt < \varepsilon \quad \text{whenever} \quad x \in K_0 \quad \text{and} \quad \text{mes } T < \delta. \tag{2.12}$$

In particular, (2.12) implies that

$$|x(t_1) - x(t_2)| \leqq \int_{t_1}^{t_2} |\dot{x}(t)|\, dt < \varepsilon \quad \text{for all } x \in K_0 \tag{2.13}$$

if $0 \leqq t_1 < t_2 \leqq 1, t_2 - t_1 < \delta$. The functions in K_0 are thus equicontinuous on $[0, 1]$, as well as, of course, uniformly bounded pointwise. The strong compactness of K_0 in $\mathscr{C}$ then follows from the theorem of Ascoli–Arzela.

Now let K be a convex subset of $\mathscr{A}$ which is closed and locally compact with respect to $w_{\mathscr{A}}$, the weak topology on $\mathscr{A}$. Let $n_{\mathscr{C}}$ and $w_{\mathscr{C}}$ denote the norm topology and weak topology on $\mathscr{C}$, respectively. Let x be an arbitrary point of K, and let U_1 be a $w_{\mathscr{A}}$-closed, convex $w_{\mathscr{A}}$-neighborhood of x such that $K \cap U_1$ is $w_{\mathscr{A}}$-compact. Certainly $K \cap U_1$ is then also $n_{\mathscr{C}}$-compact and $w_{\mathscr{C}}$-compact, as just pointed out. Thus, to prove that K is locally compact in $n_{\mathscr{C}}$ and $w_{\mathscr{C}}$ it suffices to demonstrate the existence of a $w_{\mathscr{C}}$-closed $w_{\mathscr{C}}$-neighborhood U_2 of x in $\mathscr{C}$ such that

$$K \cap U_2 \subset K \cap U_1. \tag{2.14}$$

For notational simplicity, we can suppose that $x = 0$. Let

$$U_1' = \tfrac{1}{2}U_1 \subset w_{\mathscr{A}}\text{-int } U_1, \tag{2.15}$$

$$W = K \cap [U_1 \setminus w_{\mathscr{A}}\text{-int } U_1']. \tag{2.16}$$

Then W is a $w_{\mathscr{A}}$-closed subset of $K \cap U_1$, hence $w_{\mathscr{A}}$-compact and consequently $w_{\mathscr{C}}$-compact. Also, $0 \notin W$. Therefore it is possible to select a $w_{\mathscr{C}}$-closed, convex $w_{\mathscr{C}}$-neighborhood U_2 of 0 such that $W \cap U_2 = \varnothing$. Then (2.14) must hold, for if U_2 contained a point y in K but not in U_1, the line segment joining 0 and y would contain points of $U_1 \setminus \text{int } U_1'$. Such points would lie in $W \cap U_2$ by convexity, contradicting $W \cap U_2 = \varnothing$.

It remains to show that K is also $n_{\mathscr{C}}$-closed, and therefore, by convexity, $w_{\mathscr{C}}$-closed. Let α be any positive real number, and let

$$K_\alpha = \{x \in K \mid \|x\|_{\mathscr{C}} \leqq \alpha\}. \tag{2.17}$$

Then K_α is a $w_{\mathscr{A}}$-closed subset of K, hence $w_{\mathscr{A}}$-locally compact. Moreover, K_α is convex and contains no half-lines. Therefore K_α is actually $w_{\mathscr{A}}$-compact [2]. It follows that K_α is also $n_{\mathscr{C}}$-compact and in particular $n_{\mathscr{C}}$-closed. Since this is true for arbitrary α, K itself is $n_{\mathscr{C}}$-closed.

Remark. The converse of Lemma 1 fails, at least without convexity, since a sequence in $\mathscr{A}$ converging in the $\mathscr{C}$-norm need not even be bounded as a subset of $\mathscr{A}$, much less weakly compact. Thus the properties of F asserted by Theorem 1 relative to the weak topology on $\mathscr{A}$ are considerably stronger than the properties asserted relative to the weak or strong topologies on $\mathscr{C}$.

3. Existence of optimal arcs. By a *feasible arc* for our control problem, we mean an $x \in S$ such that $F(x) < +\infty$. An *optimal arc* is a feasible arc for which the infimum of F over S is attained. Theorem 1 implies that the convex sets

$$\{x \in S | F(x) \leqq \alpha\}, \quad \alpha \text{ real}, \tag{3.1}$$

are not only closed, but locally compact in the weak topologies of $\mathscr{A}$ and $\mathscr{C}$ and the strong topology of $\mathscr{C}$. Thus any slight additional condition which ensures that these sets are actually compact (and not all empty) is enough to give us a theorem on the existence of optimal arcs.

It is well known that *a locally compact convex set is compact if and only if it does not contain any half-lines* [2]. We have already made use of this fact in proving Lemma 1. Thus the sets (3.1) are compact (in all the topologies mentioned) if S does not contain any half-line along which F is (finitely) bounded above (and hence, by convexity, "nonincreasing"). Therefore, the latter condition guarantees the existence of an optimal arc, provided there is at least one feasible arc. This criterion for existence is geometrically appealing, but not specific enough for most applications.

We proceed to formulate the half-line condition equivalently as an assumption on the sets $X(t)$ and functions l and $f(t, \cdot, \cdot)$ appearing in the control problem. For each t we denote by $\hat{X}(t)$ the *recession cone* (asymptotic cone) of $X(t)$:

$$\hat{X}(t) = \{z \in R^n | X(t) + z \subset X(t)\}. \tag{3.2}$$

We denote by $\hat{l}$ the *recession function* of l. Thus

$$\hat{l}(c_0, c_1) = \lim_{\lambda \to +\infty} l(\bar{c}_0 + \lambda c_0, \bar{c}_1 + \lambda c_1)/\lambda, \tag{3.3}$$

where $(\bar{c}_0, \bar{c}_1)$ is any element of the set E in (1.5). (The limit is independent of the particular choice of $(\bar{c}_0, \bar{c}_1)$ [13, p. 66].) Similarly, we let $\hat{f}(t, \cdot, \cdot)$ denote the recession function of $f(t, \cdot, \cdot)$.

BOUNDEDNESS ASSUMPTION. *There does not exist a nonzero arc $z \in \mathscr{A}$ such that*

$$z(t) \in \hat{X}(t) \quad \textit{for every } t, \tag{3.4}$$

$$\int_0^1 \hat{f}(t, z(t), \dot{z}(t) - A(t)z(t))\, dt + \hat{l}(z(0), z(1)) \leqq 0. \tag{3.5}$$

The integral in (3.5) is well-defined under our previous integrability assumption, and in fact it is a lower semicontinuous convex function of $z \in \mathscr{A}$ [12, Prop. 6].

THEOREM 2. *Suppose there is at least one feasible arc, and the integrability assumption is satisfied. Then the boundedness assumption is necessary for any nonempty level set of the form* (3.1) *to be weakly compact in $\mathscr{A}$ or strongly compact in $\mathscr{C}$, and it is sufficient for them all to be both weakly compact in $\mathscr{A}$ and strongly compact in $\mathscr{C}$. In particular, the boundedness assumption ensures the existence of an optimal arc. Indeed, every minimizing sequence of feasible arcs has a subsequence which converges to an optimal arc, not only in the uniform norm $\| \cdot \|_{\mathscr{C}}$, but also in the weak topology of $\mathscr{A}$.*

Proof. Let $\hat{S}$ denote the set of all arcs $z \in \mathscr{A}$ satisfying (3.4). Let $\hat{F}(z)$ denote the left side of (3.5). Then $\hat{S}$ is the recession cone of S and, as shown in [12, Prop. 6], $\hat{F}$

is the recession function of F, since for the function L in (2.9) we have

$$\hat{L}(t, z, \dot{z}) = \hat{f}(t, z, \dot{z} - A(t)z). \tag{3.6}$$

Therefore, a half-line

$$\{x + \lambda z | 0 \leqq \lambda < +\infty\}, \quad z \neq 0,$$

is contained in a set of the form (3.1) if and only if x belongs to this set and $z \in \hat{S}$, $\hat{F}(z) \leqq 0$. The conclusion of the theorem is immediate from this and Theorem 1.

COROLLARY 1. *Under the integrability assumption, an optimal arc x exists if a feasible arc exists and the sets $X(t)$ are all bounded (since then the boundedness assumption is satisfied).*

Proof. In this case (3.4) holds only for the zero arc, because $\hat{X}(t) = \{0\}$ for every t.

It is interesting to note that the boundedness of every $X(t)$ in Corollary 1 does not necessarily entail the boundedness of S, even in the norm $\| \cdot \|$, since no assumption has been made on the behavior of $X(t)$ with respect to t.

COROLLARY 2. *Under the integrability assumption, every (convex) set of the form*

$$\{x \in S | F(x) \leqq \alpha, \|x\|_{\mathscr{C}} \leqq \beta\}, \quad \alpha \text{ and } \beta \text{ real}, \tag{3.7}$$

is weakly compact in $\mathscr{A}$ and strongly compact in $\mathscr{C}$.

Proof. Apply Corollary 1 to

$$X_\beta(t) = \{x \in X(t) | |x| \leqq \beta\}. \tag{3.8}$$

COROLLARY 3. *Suppose that*

$$f(t, x, u) = f_0(t, x) + f_1(t, u), \tag{3.9}$$

$$l(x(0), x(1)) = l_0(x(0)) + l_1(x(1)), \tag{3.10}$$

and denote the recession functions of $f_i(t, \cdot, \cdot)$ and l_i by $\hat{f}_i(t, \cdot, \cdot)$ and $\hat{l}_i$, respectively. Under the integrability assumption, an optimal arc x exists if a feasible arc exists and every solution $z \in \mathscr{A}$ to the differential equation

$$\dot{z}(t) = A(t)z(t) \quad \text{a.e.} \tag{3.11}$$

satisfying (3.4) *and*

$$\int_0^1 \hat{f}_0(t, z(t))\, dt + \hat{l}_0(z(0)) + \hat{l}_1(z(1)) \leqq 0 \tag{3.12}$$

has $z(t) = 0$ for at least one $t \in [0, 1]$. (The boundedness assumption is satisfied in this case.)

Proof. If f and l have the structure in (3.9) and (3.10), their recession functions also have this structure:

$$\hat{f}(t, z, y) = \hat{f}_0(t, z) + \hat{f}_1(t, y), \tag{3.13}$$

$$\hat{l}(z(0), z(1)) = \hat{l}_0(z(0)) + \hat{l}_1(z(1)). \tag{3.14}$$

Furthermore, the integrability assumption implies that

$$\sup_{u \in R^n} \{u \cdot p - f_1(t, u)\} < +\infty \quad \text{for all } p \in R^n, \tag{3.15}$$

and consequently [13, p. 116] that

$$\hat{f}_1(t, y) = \delta_0(y) = \begin{cases} 0 & \text{if } y = 0, \\ +\infty & \text{if } y \neq 0. \end{cases} \tag{3.16}$$

Thus in this case condition (3.5) is equivalent to (3.11) and (3.12). The result now follows from Theorem 2 and the fact that a solution to (3.11) which vanishes for some $t \in [0, 1]$ must be the zero arc.

Remark. A simple but common case where the condition in Corollary 3 is satisfied occurs when one of the (convex) endpoint sets

$$E_i = \{c \in R^n | l_i(c) < +\infty\}, \quad i = 0, 1, \tag{3.17}$$

is bounded (so that $\hat{l}_i$ is finite only at 0), or one of the sets $X(t)$ is bounded (so that $\hat{X}(t) = \{0\}$). A more general result resembling Corollary 3 can be derived from [12, Cor. 3 to Thm. 3].

4. The dual control problem. The question of necessary and sufficient conditions for the optimality of an arc x is closely tied in with duality, a topic of interest in its own right. In this section we describe the basic duality briefly, to set the stage for later developments.

The dual control problem is:

$$\text{Minimize} \quad \int_0^1 g(t, p(t), w(t))\, dt + m(p(0), p(1)) \tag{4.1}$$

subject to

$$\dot{p}(t) = -A^*(t)p(t) + w(t) \quad \text{for almost every } t, \tag{4.2}$$

where $A^*(t)$ is the transpose of $A(t)$ and

$$g(t, p, w) = \sup \{u \cdot p + x \cdot w - f(t, x, u) | u \in U(t, x), x \in X(t)\}, \tag{4.3}$$

$$m(d_0, d_1) = \sup \{c_0 \cdot d_0 - c_1 \cdot d_1 - l(c_0, c_1) | c_0 \in R^n, c_1 \in R^n\}. \tag{4.4}$$

In terms of the Bolza functional

$$G(p) = \int_0^1 g(t, p(t), \dot{p}(t) + A^*(t)p(t))\, dt + m(p(0), p(1)), \tag{4.5}$$

we can express this problem as:

$$\text{Minimize} \quad G(p) \quad \text{over all} \quad p \in \mathscr{A}. \tag{4.6}$$

Before discussing the circumstances under which the integral in the Bolza functional G is well-defined, we remark that the dual problem, like the original problem, involves implicit constraints on the controls and endpoints. The implicit dual control set is

$$W(t, p) = \{w \in R^n | g(t, p, w) < +\infty\}, \tag{4.7}$$

while the endpoint pair $(p(0), p(1))$ must belong to the set where the function m is finite-valued. However, there are no *state* constraints in the dual problem, even implicit ones. In particular,

$$W(t,p) \neq \varnothing \quad \text{for every } t \in [0,1] \text{ and } p \in R^n. \tag{4.8}$$

This is evident from the following result, which may also be helpful in calculating $g(t,p,w)$ in specific cases.

LEMMA 2. *In terms of the function h in* (2.6) *and the functions*

$$s(t,w) = \sup\{x \cdot w | x \in X(t)\}, \tag{4.9}$$

$$k(t,p,w) = \sup\{x \cdot w + h(t,x,p) | x \in R^n\}, \tag{4.10}$$

one has, under the integrability assumption,

$$\begin{aligned} g(t,p,w) &= \sup\{x \cdot w + h(t,x,p) | x \in X(t)\} \\ &= \min\{k(t,p,w-z) + s(t,z) | z \in R^n\}. \end{aligned} \tag{4.11}$$

Furthermore, for every measurable, essentially bounded function $p:[0,1] \to R^n$ it is possible to find a summable function $w:[0,1] \to R^n$ and a summable function $\alpha:[0,1] \to R^1$ such that

$$g(t,p(t),w(t)) \leqq \alpha(t) \quad \textit{for all } t. \tag{4.12}$$

Proof. The first equality in (4.11) is immediate from the definitions of g and h. The second equality results from a basic theorem of convex analysis expressing the conjugate of the sum of two convex functions in terms of infimal convolution of the conjugate functions [13, Thm. 16.4]: according to the first equality in (4.11), $g(t,p,\cdot)$ is the conjugate of the sum of $-h(t,\cdot,p)$ and the indicator of the set $X(t)$, while, by definition, $k(t,p,\cdot)$ is the conjugate of $-h(t,\cdot,p)$, and $s(t,\cdot)$ is the conjugate of the indicator of $X(t)$. The theorem in question is applicable, because $-h(t,\cdot,p)$ is a finite function under the integrability assumption. (The convexity of $-h(t,x,p)$ in x has already been noted in § 2.) In the notation and terminology of [10], [12], the function

$$M(t,p,r) = k(t,p,r + A^*(t)p) \tag{4.13}$$

is the Lagrangian dual to the function L in (2.9). We have shown in the proof of Theorem 1 that L satisfies conditions (A), (B), (C_0) and (D_0) of [12] under the integrability assumption, and this implies that M satisfies the same conditions [10, Thm. 2], [12, § 1]. Then for every measurable, essentially bounded function $p:[0,1] \to R^n$ it is possible to find a summable function $r:[0,1] \to R^n$ and a summable function $\alpha:[0,1] \to R^1$ such that

$$M(t,p(t),r(t)) \leqq \alpha(t) \quad \text{for every } t. \tag{4.14}$$

Setting $w(t) = r(t) + A^*(t)p(t)$, we obtain a summable function w for which (4.12) holds, since $g \leqq k$. The lemma is thereby proved.

We now introduce a further condition from which it will be deduced, in particular, that the integral in (4.5) is well-defined.

INTERIORITY ASSUMPTION. *The multifunction $X:t \to X(t)$ satisfies*

$$\operatorname{int} X(t) \neq \varnothing \quad \textit{for every } t \tag{4.15}$$

and

$$\{(t, x) | x \in \operatorname{int} X(t)\} = \operatorname{int} \operatorname{cl} \{(t, x) | x \in X(t)\}. \tag{4.16}$$

For an equivalent form of this assumption, see [14, Lemma 2].

LEMMA 3. *Under the interiority assumption, the constraint set S has a nonempty interior in $\mathscr{A}$ consisting of the functions x such that*

$$x(t) \in \operatorname{int} X(t) \quad \textit{for every } t, \tag{4.17}$$

and this is the same as the interior of S relative to the norm $\| \cdot \|_{\mathscr{C}}$. Furthermore, any $x \in \mathscr{A}$ satisfying

$$x(t) \in X(t) \quad \textit{for almost every } t \tag{4.18}$$

actually satisfies (1.3). *Thus any $x \in \mathscr{A}$ satisfying* (4.18) *belongs to S.*

Proof. If $\mathscr{A}$ were replaced by $\mathscr{C}$ here (and in the definition of S), this would be the special case of [14, Thm. 5 and Lemma 2], where the $D(t)$ and $f(t, \cdot)$ in the notation of the latter results are taken to be $X(t)$ and the indicator of $X(t)$, respectively. The lemma is also valid in the present form, because $\mathscr{A}$ is dense in $\mathscr{C}$.

THEOREM 3. *Under the assumptions of integrability and interiority, G in* (4.5) *is a well-defined, lower semicontinuous, convex functional from $\mathscr{A}$ to $R^1 \cup \{+\infty\}$. If the boundedness assumption also holds, then*

$$-\min_{x \in S} F(x) = \inf_{p \in \mathscr{A}} G(p) < +\infty. \tag{4.19}$$

Proof. We deduce this from the main theorem of [12]. Let L be as in (2.9), and let

$$L_0(t, x, v) = \begin{cases} L(t, x, v) & \text{if } x \in X(t), \\ +\infty & \text{if } x \notin X(t). \end{cases} \tag{4.20}$$

The measurability properties of L_0, to be described in a moment, ensure that the Bolza functional

$$F_0(x) = \int_0^1 L_0(t, x(t), \dot{x}(t))\, dt + l(x(0), x(1)) \tag{4.21}$$

is well-defined. The given control problem can be regarded as that of minimizing F_0 over all of $\mathscr{A}$, since from the second assertion of Lemma 3 we have

$$F_0(x) = \begin{cases} F(x) & \text{if } x \in S, \\ +\infty & \text{if } x \notin S. \end{cases} \tag{4.22}$$

The corresponding dual problem, in the terminology of [10] and [12], is that of minimizing

$$\int_0^1 M_0(t, p(t), \dot{p}(t))\, dt + m(p(0), p(1)) \tag{4.23}$$

over all $p \in \mathscr{A}$, where

$$M_0(t, p, r) = \sup_{x, v} \{x \cdot r + v \cdot p - L_0(t, x, v)\}, \tag{4.24}$$

and m is given by (4.4). It is easily seen from the definitions that

$$M_0(t, p, r) = g(t, p, r + A^*(t)p), \tag{4.25}$$

so that this dual problem is indeed the problem presented here as the dual.

We shall verify that conditions (A), (B), (C_0) and (D) of [12] hold for L_0. The assertions about the nature of G will then be valid by [10, Thm. 2]. The duality relation (4.19) will follow from Theorem 1(a) of [12], since our boundedness assumption is equivalent by [12, Cor. 1 of Thm. 3] to the condition of dual attainability in the hypothesis of that theorem.

Since $X(t)$ is a closed, convex set, the function $L_0(t, \cdot, \cdot)$ is lower semicontinuous and convex on $R^n \times R^n$. Thus condition (A) of [12] is satisfied by L_0, provided that $L_0(t, \cdot, \cdot)$ is not identically $+\infty$ for any t. The latter will be seen in a moment, in connection with condition (D).

We have already verified in the proof of Theorem 1 that conditions (A), (B), (C_0) and (D_0) of [12] are satisfied by the function L, and in the case of (B) this means that L is measurable with respect to the σ-field in $[0, 1] \times R^{2n}$ generated by products of the Lebesgue sets in $[0, 1]$ and Borel sets in R^{2n}. The corresponding measurability property of L_0 then follows from the measurability of the set

$$\{(t, x, v) | x \in X(t)\} \subset [0, 1] \times R^{2n}, \tag{4.26}$$

which is implied by our interiority assumption. Indeed, since $X(t)$ is a closed, convex set with nonempty interior, we have $x \in X(t)$ if and only if $x + B_j$ meets int $X(t)$ for every natural number j, where B_j is the open ball of radius $1/j$ and center 0 in R^n. Furthermore, the openness of the set on the left in (4.16) implies the openness of the set of all (t, x) such that $x + B_j$ meets int $X(t)$. Thus the set (4.26) can be expressed as the intersection of a countable collection of open sets and in particular is Borel measurable.

The fact that L satisfies (C_0) trivially implies that L_0 satisfies (C_0), since $L_0 \geqq L$. As for condition (D), we must demonstrate the existence of a bounded, measurable function $x:[0, 1] \to R^n$, a summable function $v:[0, 1] \to R^n$, and a summable function $\alpha:[0, 1] \to R^1$ such that

$$L_0(t, x(t), v(t)) \leqq \alpha(t) \quad \text{for every } t, \tag{4.27}$$

or, in other words, such that (1.3) holds and

$$L(t, x(t), v(t)) \leqq \alpha(t) \quad \text{for every } t. \tag{4.28}$$

Since L is known to satisfy condition (D_0), corresponding functions v and α satisfying (4.28) exist for any bounded, measurable function x [12, Prop. 3]. Therefore, it is enough to show there is at least one bounded, measurable function x for which (1.3) holds. In fact, Lemma 3 asserts the existence of such a function x belonging to $\mathscr{A}$.

COROLLARY. *Under the assumptions of integrability, boundedness, and interiority, a feasible arc x exists in the original control problem if and only if, in the dual problem,*

$$\inf_{p \in \mathscr{A}} G(p) > -\infty. \tag{4.29}$$

Proof. A feasible arc exists if and only if the infimum of F over S in (4.19) is not $+\infty$.

5. Optimality conditions without state constraints. In the case where $X(t) = R^n$ for every t, it is possible to derive from the results in [10] and [12] a companion theorem to Theorem 3, furnishing conditions that are necessary and sufficient for the optimality of an arc x. We carry this out as a step in the derivation of more general optimality conditions in § 6 for the case where (nontrivial) state constraints are present. The conditions we formulate at this stage do in fact have some bearing on state constraints, and they help to clarify the relationship between the given control problem and its dual.

As usual, we denote by $\partial\varphi(z)$ the (closed, convex) set of all *subgradients* of a convex function φ on R^n at the point z, that is, the set of all vectors y such that

$$\varphi(z') \geqq \varphi(z) + y' \cdot (z' - z) \quad \text{for all } z' \in R^n. \tag{5.1}$$

Subgradients of a concave function are defined analogously, with the reversed inequality. The relationship between subgradients and ordinary gradients is discussed at length in [13, § 25]. For the function h in (2.6), we denote by $\partial_p h(t, x, p)$ the set of subgradients of the convex function $h(t, x, \cdot)$ at p and by $\partial_x h(t, x, p)$ the set of subgradients of the concave function $h(t, \cdot, p)$ at x. These sets are described further in Lemma 4 below.

We denote by $N(t, x)$ the *cone of normals* [13] to $X(t)$ at the point x. This is the closed, convex cone defined by

$$N(t, x) = \begin{cases} \{w \mid w \cdot (z - x) \leqq 0 \text{ for all } z \in X(t)\} & \text{if } x \in X(t), \\ \varnothing & \text{if } x \notin X(t). \end{cases} \tag{5.2}$$

The conditions to be analyzed may now be stated.

OPTIMALITY CONDITIONS. *The functions $x \in \mathscr{A}$ and $p \in \mathscr{A}$ satisfy*

$$u(t) \in \partial_p h(t, x(t), p(t)) \quad a.e., \tag{5.3}$$

$$w(t) \in -\partial_x h(t, x(t), p(t)) + N(t, x(t)) \quad a.e., \tag{5.4}$$

where

$$\dot{x}(t) = A(t)x(t) + u(t) \quad \text{and} \quad \dot{p}(t) = -A^*(t)p(t) + w(t) \quad a.e. \tag{5.5}$$

Furthermore,

$$(p(0), -p(1)) \in \partial l(x(0), x(1)) \quad (\textit{transversality}). \tag{5.6}$$

The state constraint (1.3) on x is embedded in (5.4) by virtue of (5.2) and Lemma 3, if the interiority condition holds.

The relationship between these optimality conditions and the familiar "maximum principle" is made clearer by the following result. (Note the *normality*: the multiplier of the cost function is taken to be -1.)

LEMMA 4. *Let the integrability assumption be satisfied. Then the set $\partial_p h(t, x, p)$ consists of the control vectors $u \in U(t, x)$ for which the supremum of $u \cdot p - f(t, x, u)$ is attained. On the other hand, in terms of the dual cost function g and dual control set W, the set $-\partial_x h(t, x, p) + N(t, x)$ consists of the dual control vectors $w \in W(t, p)$ for which the supremum of $x \cdot w - g(t, p, w)$ is attained.*

Moreover, if the function H_0 is defined by

$$H_0(t,x,p) = \begin{cases} h(t,x,p) + p \cdot A(t)x & \text{if } x \in X(t), \\ -\infty & \text{if } x \notin X(t), \end{cases} \tag{5.7}$$

the optimality conditions, except for transversality, can be expressed in the Hamiltonian form

$$\dot{x} \in \partial_p H_0(t,x,p) \quad \text{and} \quad -\dot{p} \in \partial_x H_0(t,x,p) \quad \text{a.e.} \tag{5.8}$$

Proof. The assertion concerning $\partial_p h(t,x,p)$ is immediate from the fact that $h(t,x,\cdot)$ is by definition the conjugate of the convex function $f(t,x,\cdot)$, whose effective domain is $U(t,x)$ (see [13, Thm. 23.5]). On the other hand, let

$$h_0(t,x,p) = \begin{cases} h(t,x,p) & \text{if } x \in X(t), \\ -\infty & \text{if } x \notin X(t). \end{cases} \tag{5.9}$$

Then $g(t,p,\cdot)$ is by (4.11) the conjugate of the convex function $-h_0(t,\cdot,p)$, and the effective domain of $g(t,p,\cdot)$ is $W(t,p)$. If the integrability assumption is satisfied, so that $h(t,x,p)$ is finite and hence continuous as a concave function of x, $-h_0(t,\cdot,p)$ is lower semicontinuous ($X(t)$ being closed) and consequently is in turn the conjugate of $g(t,p,\cdot)$. Thus, by the same reasoning as for $\partial_p h(t,x,p)$, the set $-\partial h_0(t,x,p)$ consists of the vectors $w \in W(t,p)$ for which the supremum of $x \cdot w - g(t,p,w)$ is attained. We note now that, since $-h_0(t,\cdot,p)$ is the sum of the finite convex function $-h(t,\cdot,p)$ and the indicator of the nonempty convex set $X(t)$, we have

$$-\partial h_0(t,x,p) = -\partial h(t,x,p) + N(t,x). \tag{5.10}$$

This is a special case of a basic formula for subdifferentiation (see [13, p. 215 and Thm. 23.8]). The same formula also yields the Hamiltonian form of the optimality conditions.

The generalized Hamiltonian equations (5.8) have been studied in [10] and [11]. They are very similar to ordinary differential equations, despite the "multivaluedness". As a matter of fact, if the integrability and interiority assumptions are satisfied, the subgradient sets in (5.8) reduce to single elements, the ordinary gradients of H_0, for almost all choices of (t,x,p) such that $x \in \operatorname{int} X(t)$. This result may be deduced from [13, Thm. 35.9].

The next lemma explains the fundamental connection between the optimality conditions and the dual control problem (cf. Thm. 3).

LEMMA 5. *Let the assumptions of integrability and interiority hold. Then, in order that $x \in S$ and $p \in \mathscr{A}$ be arcs such that*

$$-\min_S F = -F(x) = G(p) = \min_{\mathscr{A}} G, \tag{5.11}$$

it is necessary and sufficient that the preceding optimality conditions be satisfied.

Proof. As we have observed at the beginning of the proof of Theorem 3, the given control problem corresponds under our assumptions to the problem of minimizing the Bolza functional F_0 in (4.21) over all of $\mathscr{A}$. Thus this is just the special case of [10, Thm. 5] corresponding to l and the function L in (4.20).

COROLLARY. *Let the assumptions of integrability and interiority hold, and let $x \in \mathscr{A}$. In order that x be an optimal arc in the given control problem, it is sufficient*

that there exist an arc $p \in \mathcal{A}$ such that the optimality conditions are satisfied. These conditions are also necessary, if the boundedness assumption holds and the infimum of G in the dual problem is attained.

Proof. The sufficiency is obvious from the lemma, and the necessity is a consequence of Theorem 3.

The preceding corollary indicates the need of some criterion for the attainment of the infimum in the dual control problem. This is where state constraints in the original problem cause difficulties. A dual existence theorem ought to involve some Nagumo–Tonelli-type growth condition on the convex functions $g(t, p, \cdot)$: at the very least,

$$\lim_{\lambda \to +\infty} g(t, p, \bar{w} + \lambda w)/\lambda = +\infty \tag{5.12}$$

for every $\bar{w}$ and nonzero w. However, this growth property is not present if $X(t) \neq R^n$. Indeed, since $g(t, p, \cdot)$ is the conjugate of the convex function $-h_0(t, \cdot, p)$, where h_0 is given by (5.9), the property holds if and only if $h_0(t, x, p)$ is finite as a function of x [13, Cor. 13.3.1], and this is not true if $X(t) \neq R^n$.

These observations show that we cannot hope to prove the *necessity* of the above optimality conditions in the case of nontrivial state constraints. However, this is not surprising. It is well known that state constraints can cause "jumps" in the costate variables $p(t)$, yet the optimality conditions, as we have stated them, do not allow for such discontinuities. There is a dual interpretation of the situation: the reason we cannot prove necessity is that the dual problem is formulated too narrowly for solutions to exist. If the dual cost function g does not have the growth properties which guarantee that the Bolza functional G attains its infimum over $\mathcal{A}$, then perhaps G can be extended to a larger space of functions p for which we do have attainment. We shall reconcile these two interpretations in § 7. For the present, we state a result applicable to the case where $X(t) = R^n$.

ATTAINABILITY ASSUMPTION. *The relative interiors of the convex sets D and E in $R^n \times R^n$ have a nonempty intersection, where E is the set of all "admissible" endpoint pairs defined in* (1.5), *and D is the set of all pairs $(x(0), x(1))$ arising from arcs $x \in \mathcal{A}$ such that $\dot{x}(t) = A(t)x(t) + u(t)$ with $u(t) \in U(t, x(t))$ for almost every t.*

The relative interior of a convex set is the interior relative to the affine hull of the set [13, § 6]. Note that one does at least have $D \cap E \neq \varnothing$ if a feasible arc x exists. Thus, if there are not state constraints, the attainability assumption says that "feasible arcs exist, and not just marginally."

THEOREM 4 [12]. *Suppose that $X(t) = R^n$ for every $t \in [0, 1]$, so that S is all of $\mathcal{A}$ and the normal cone $N(t, x)$ is just $\{0\}$ for every $(t, x) \in [0, 1] \times R^n$. Let the assumptions of integrability and attainability hold. Then*

$$-\inf_{\mathcal{A}} F = \min_{\mathcal{A}} G > -\infty. \tag{5.13}$$

In order that $x \in \mathcal{A}$ be an optimal arc in the original control problem, it is necessary and sufficient that there exist an arc $p \in \mathcal{A}$ for which the optimality conditions hold. The arcs p obtained in this way are precisely the optimal arcs in the dual problem.

Proof. This combines Lemma 4 with [12, Thm. 1(b)] for the function L in (2.9). (Condition (D_0) of [12] holds for L, as observed in the proof of Theorem 1 above. Furthermore, the set C_L in [12] has the same relative interior as the set D

here, because C_L has the same relative interior as

$$\left\{(x(0), x(1)) | x \in \mathscr{A}, \int_0^1 L(t, x(t), \dot{x}(t))\, dt < \infty\right\} \tag{5.14}$$

according to [12, Cor. 4 of Thm. 3], and D lies between these two sets.)

6. Generalized optimality conditions for state constraints. We proceed to reduce the general case to the case of no state constraints by means of an abstract multiplier principle.

Every continuous linear functional on the Banach space C corresponds, of course, to an R^n-valued measure on $[0, 1]$, which can be expressed as dp for a certain function $p:[0, 1] \to R^n$ of *bounded variation*. We denote by $\mathscr{B}$ the space of all such functions under the norm

$$\|p\|_{\mathscr{B}} = |p(0)| + \|dp\|. \tag{6.1}$$

Thus $\mathscr{B}$ is a Banach space isomorphic to $R^n \times C^*$, and $\mathscr{A}$ is isometrically embedded in $\mathscr{B}$. It should be said that $\mathscr{B}$ actually consists of equivalence classes: we identify two functions p and q if $p(t^-) = q(t^-)$ and $p(t^+) = q(t^+)$ at all the (countably many) places where these functions have jumps, since then $dp = dq$. In this context we regard $p(0)$ as $p(0^-)$ and $p(1)$ as $p(1^+)$.

STRICT FEASIBILITY ASSUMPTION. *There is at least one arc $x \in \mathscr{A}$ satisfying $F(x) < +\infty$ such that $x(t) \in \operatorname{int} X(t)$ for every t.*

LEMMA 6. *Let the assumptions of integrability, interiority and strict feasibility hold. Then an arc $x \in \mathscr{A}$ is optimal in the original control problem (that is, F attains its minimum over S at x), if and only if there is a function $p_0 \in \mathscr{B}$ for which the linear functional*

$$\begin{aligned} \Lambda(z) &= -\int_0^1 p_0(t) \cdot \dot{z}(t)\, dt - p_0(0) \cdot z(0) + p_0(1) \cdot z(1) \\ &= \int_0^1 z(t)\, dp_0(t), \quad z \in \mathscr{A}, \end{aligned} \tag{6.2}$$

has the following properties:

(a) *Λ attains its maximum over S at x;*

(b) *$F + \Lambda$ attains its minimum over $\mathscr{A}$ at x.*

Proof. The sufficiency is immediate from the fact that (a) and (b) imply, for arbitrary $z \in S$,

$$F(z) = (F + \Lambda)(z) - \Lambda(z) \geqq (F + \Lambda)(x) - \Lambda(x) = F(x). \tag{6.3}$$

To prove the necessity, let x denote any optimal arc. In the space $\mathscr{A} \times R^1$, we consider two convex sets, the epigraph of F (i.e., the set of all pairs (z, α) such that $\alpha \geqq F(z)$) and $S \times (-\infty, F(x)]$. The latter has a nonempty interior which does not meet the former, so that the two sets can be separated by a closed hyperplane. The hyperplane cannot be "vertical," because of our strict feasibility assumption, and therefore it is the graph of a certain continuous linear functional Λ on $\mathscr{A}$. Since both sets contain the point $(x, F(x))$, properties (a) and (b) hold for Λ. Furthermore, since S has a nonempty interior relative to the norm $\|\cdot\|_{\mathscr{C}}$ (Lemma 3), Λ must

actually be continuous relative to $\|\cdot\|_{\mathscr{C}}$, due to (a), and hence Λ can be represented as in (6.2). (For the integration by parts, see [4].)

The virtue of Lemma 6 is that the minimization in (b) corresponds to a control problem without state constraints. Namely, one has

$$(F + \Lambda)(x) = \int_0^1 f_1(t, x(t), \dot{x}(t) - A(t)x(t))\,dt + l_1(x(0), x(1)), \tag{6.4}$$

where

$$f_1(t, x, u) = f(t, x, u) - p_0(t)\cdot(A(t)x + u), \tag{6.5}$$

$$l_1(x(0), x(1)) = l(x(0), x(1)) - p_0(0)\cdot x(0) + p_0(1)\cdot x(1). \tag{6.6}$$

The functions f_1 and l_1 satisfy the same assumptions as f and l. Thus we can use Theorem 4 of the preceding section to characterize (b).

At the same time, the situation in (a) of Lemma 6 can be characterized by results in [14]. These results make use of the following concept.

Let $K(t)$ denote a convex cone in R^n (containing the origin) for each $t \in [0, 1]$. An R^n-valued measure μ on $[0, 1]$ is said to be *$K(t)$-valued* if the Radon–Nikodym derivative $d\mu/d\theta$ satisfies

$$\frac{d\mu}{d\theta}(t) \in K(t), \quad \theta\text{-a.e.}, \tag{6.7}$$

where θ is any positive measure on $[0, 1]$ with respect to which μ is absolutely continuous. (This property is independent of the particular θ.)

LEMMA 7. *Under the interiority assumption, a functional of the form* (6.2) *attains its maximum over S at x if and only if* ($x \in S$ *and*) *the measure dp_0 is $N(t, x(t))$-valued* (*where $N(t, x(t))$ is the cone of normals to $X(t)$ at $x(t)$*).

Proof. If $\mathscr{A}$ were replaced by $\mathscr{C}$ in the definition of S this would be Corollary 6A of [14], since our interiority assumption on the multifunction $X : t \to X(t)$ implies lower semicontinuity [14, Lemma 2]. The result follows in the present case because Λ is continuous in the norm $\|\cdot\|_{\mathscr{C}}$, and S as a subset of $\mathscr{A}$ is dense in the corresponding subset of $\mathscr{C}$ (Lemma 3).

Our main result on necessary and sufficient conditions, Theorem 5 below, concerns the following conditions.

GENERALIZED OPTIMALITY CONDITIONS. *These are the same as the optimality conditions in* § 5, *except that $p \in \mathscr{B}$ rather than $p \in \mathscr{A}$, and*

$$\textit{the singular part of } dp \textit{ is } N(t, x(t))\textit{-valued}. \tag{6.8}$$

Of course, if p is of bounded variation, the derivative $\dot{p}(t)$ in condition (5.5) does exist for almost every t, although it is not necessarily true that p is the integral of $\dot{p}$. The singular part of the measure dp may be regarded as the "singular dual control contribution," in the sense that one has

$$dp(t) = -A^*(t)p(t)\,dt + d\mu(t), \tag{6.9}$$

where

$$d\mu(t) = w(t)\,dt + (\text{singular part}). \tag{6.10}$$

The generalized optimality conditions reduce to the previous ones if there are no state constraints, since then $N(t, x(t)) \equiv \{0\}$. (To say that the singular part of dp is $\{0\}$-valued is to say that p is absolutely continuous.) More generally, condition (6.8) implies that the singular part of dp is concentrated in the set of t values for which $x(t)$ lies on the boundary of $X(t)$. If p is discontinuous at t, we get the jump condition

$$p(t^+) - p(t^-) \in N(t, x(t)). \tag{6.11}$$

The generalized optimality conditions, except for transversality, may be regarded as the natural extension of the Hamiltonian "equations" (5.8) to allow for $p \in \mathscr{B}$, instead of just $p \in \mathscr{A}$.

THEOREM 5. *Let the assumptions of integrability, interiority, attainability and strict feasibility be satisfied. Then, in order that an arc $x \in \mathscr{A}$ be optimal in the given control problem, it is necessary and sufficient that there exist a function $p \in \mathscr{B}$ for which the generalized optimality conditions are satisfied.*

Proof. We are in the situation of Lemma 6. Property (a) of Lemma 6 has already been characterized in Lemma 7. On the other hand, we have observed that property (b) characterizes x as an optimal arc for a certain control problem without state constraints, corresponding to the functions f_1 and l_1 in (6.5) and (6.6). The integrability and attainability assumptions on f and l carry over to f_1 and l_1, as may easily be checked. Thus Theorem 4 is valid for the unconstrained problem with f_1 and l_1. The function h_1 which corresponds to f_1, as h does to f in (2.6), is

$$h_1(t, x, p) = h(t, x, p + p_0(t)) + p_0(t) \cdot A(t)x. \tag{6.12}$$

Thus the optimality conditions for the unconstrained problem, expressed in terms of a function $p_1 \in \mathscr{A}$, are:

$$\dot{x}(t) = A(t)x(t) + u(t) \quad \text{and} \quad \dot{p}_1(t) = -A^*(t)p_1(t) + w_1(t) \quad \text{a.e.}, \tag{6.13}$$

$$u(t) \in \partial_p h_1(t, x(t), p_1(t)) = \partial_p h(t, x(t), p_0(t) + p_1(t)) \quad \text{a.e.}, \tag{6.14}$$

$$w_1(t) \in -\partial_x h_1(t, x(t), p_1(t)) = -\partial_x h(t, x(t), p_0(t) + p_1(t)) - A^*(t)p_0(t) \quad \text{a.e.}, \tag{6.15}$$

$$(p_1(0), -p_1(1)) \in \partial l_1(x(0), x(1)) = \partial l(x(0), x(1)) - (p_0(0), -p_0(1)). \tag{6.16}$$

We may conclude therefore, as an intermediate step, that x is an optimal arc if and only if $x \in S$ and there exist functions $p_0 \in \mathscr{B}$ and $p_1 \in \mathscr{A}$ such that dp_0 is $N(t, x(t))$-valued and conditions (6.13) through (6.16) are satisfied.

We can write dp_0 as

$$dp_0(t) = \dot{p}_0(t)\, dt + d\mu(t), \tag{6.17}$$

where μ is a certain singular measure. Then dp_0 is $N(t, x(t))$-valued, and x belongs to S if and only if μ is $N(t, x(t))$-valued and $\dot{p}_0(t) \in N(t, x(t))$ a.e. (The latter implies that $N(t, x(t)) \neq \varnothing$ a.e., so that $x(t) \in X(t)$ a.e.; then $x \in S$ by Lemma 3.)

Suppose now that the preceding conditions are satisfied by x, p_0 and p_1, and let

$$p(t) = p_0(t) + p_1(t), \tag{6.18}$$

$$w(t) = \dot{p}_0(t) + A^*(t)p_0(t) + w_1(t). \tag{6.19}$$

Then $\dot{p}(t) = \dot{p}_0(t) + \dot{p}_1(t)$, and the singular part of dp is μ. A simple check shows that the conditions at hand reduce to the generalized optimality conditions for x and p.

Conversely, suppose that x and p satisfy the generalized optimality conditions. It would be possible to show that p can be written in the form (6.18) in such a way that the preceding intermediate conditions are satisfied. However, this approach requires a complicated measurability argument. We therefore proceed more directly, via the theory of subgradients and the fact (Lemma 4) that the optimality conditions (5.3), (5.4) and (5.5) can be expressed in the Hamiltonian form (5.8). Since H_0 is the Hamiltonian corresponding to the Lagrangian L_0 in (4.20), we can express these conditions equivalently in the Lagrangian form

$$(\dot{p}(t), p(t)) \in \partial L_0(t, x(t), \dot{x}(t)) \quad \text{a.e.}, \tag{6.20}$$

where $\partial L_0(t, \cdot, \cdot)$ denotes the set of subgradients (in $R^n \times R^n$) of the convex function $L_0(t, \cdot, \cdot)$ [10, p. 212]. As observed in the proof of Theorem 3, the given control problem amounts to minimizing the functional F_0 in (4.21) over all of $\mathscr{A}$. Thus we need only show that (6.20), (5.6) and (6.8) imply

$$F_0(z) \geqq F_0(x) \quad \text{for every } z \in \mathscr{A}. \tag{6.21}$$

Fixing $z \in \mathscr{A}$, we observe from (6.20), (5.6) and the definition of "subgradient" that

$$\begin{aligned} L_0(t, z(t), \dot{z}(t)) \geqq\ & L(t, x(t), \dot{x}(t)) + (z(t) - x(t)) \cdot \dot{p}(t) \\ & + (\dot{z}(t) - \dot{x}(t)) \cdot p(t) \quad \text{a.e.}, \end{aligned} \tag{6.22}$$

while

$$l(z(0), z(1)) \geqq l(x(0), x(1)) + (z(0) - x(0)) \cdot p(0) - (z(1) - x(1)) \cdot p(1). \tag{6.23}$$

Integrating both sides of (6.22) and adding to (6.23), we get

$$F_0(z) \geqq F_0(x) - \int_0^1 (z(t) - x(t))\, d\mu(t), \tag{6.24}$$

where μ is the singular part of dp. Unless $F_0(z) = +\infty$ (in which event $F_0(z) \geqq F_0(x)$ trivially), $z(t)$ must belong to $X(t)$ for every t (cf. Lemma 3). Then, since μ is $N(t, x(t))$-valued, the integral in (6.24) is nonpositive, so that $F_0(z) \geqq F_0(x)$.

Remark. The preceding argument shows that the generalized optimality conditions are sufficient even without the assumptions of attainability and strict feasibility, as long as at least one feasible arc exists. (The existence of a feasible arc was used in concluding from (6.21) that x is itself feasible, i.e., $F_0(x) \neq +\infty$. Actually, this assumption is unnecessary, in view of Lemma 8 in the next section.)

7. The generalized dual problem. Theorem 5 is incomplete in comparison with Theorem 4, because the meaning of the functions p that appear in the generalized optimality conditions is unexplained. Presumably such functions are generalized solutions to the dual problem. We show now that this is true in a certain precise sense.

Returning to the function s in (4.9), we define for an R^n-valued measure μ:

$$\int_0^1 s(t, d\mu(t)) = \int_0^1 s(t, (d\mu/d\theta)(t))\, d\theta(t), \tag{7.1}$$

where θ is any positive measure with respect to which μ is absolutely continuous. (Since $s(t, w)$ is positively homogeneous as a function of w, this formula is independent of the particular θ.) Under our interiority assumption, s has the measurability needed for this definition, and the integral is well-defined (possibly $+\infty$, but not $-\infty$) [14, § 4].

It will be recalled that the dual control problem in § 4 consists of minimizing the Bolza functional G in (4.5) over the space $\mathscr{B}$. We take the *generalized dual problem* to be that of minimizing over the space $\mathscr{B}$ the functional

$$\bar{G}(p) = G(p) + \int_0^1 s(t, d\mu(t)), \tag{7.2}$$

where μ is the singular part of dp. Since $s(t, 0) \equiv 0$, it is clear that $\bar{G}$ agrees with G on $\mathscr{A}$. Thus $\bar{G}$ is a certain extension of G from $\mathscr{A}$ to the larger space $\mathscr{B}$. The topological nature of this extension is described below (Theorem 6). It may be ascertained that $\bar{G}$ is again convex. (This can be deduced using (4.11) and the definitions. It is also shown by the last part of the proof of Theorem 6, which assumes only "integrability" and "interiority.")

LEMMA 8. *Let the assumptions of integrability and interiority hold. Then the generalized optimality conditions are satisfied by $x \in \mathscr{A}$ and $p \in \mathscr{B}$ if and only if $x \in S$ and*

$$-\min_S F = -F(x) = \bar{G}(p) = \min_{\mathscr{B}} \bar{G}. \tag{7.3}$$

Proof. Define L_0 as in (4.20) and M_0 as in (4.25); thus L_0 and M_0 are the Lagrangians dual to each other that correspond to the Hamiltonian H_0 given by (5.7). The Hamiltonian "equations" (5.8), which according to Lemma 4 express the optimality conditions (5.3), (5.4) and (5.5), can then [10, p. 212] be expressed as

$$L_0(t, x(t), \dot{x}(t)) + M_0(t, p(t), \dot{p}(t)) = x(t) \cdot \dot{p}(t) + x(t) \cdot p(t) \quad \text{a.e.}, \tag{7.4}$$

where for arbitrary $x \in \mathscr{A}$ and $p \in \mathscr{B}$ it would be true that

$$L_0(t, x(t), \dot{x}(t)) + M_0(t, p(t), \dot{p}(t)) \geqq x(t) \cdot \dot{p}(t) + \dot{x}(t) \cdot p(t) \quad \text{a.e.} \tag{7.5}$$

On the other hand, if θ is any positive measure on $[0, 1]$ with respect to which both Lebesgue measure on $[0, 1]$ and the singular part μ of the measure dp are absolutely continuous, we can write the conditions (6.8) and $x \in S$ as

$$r(t, x(t)) + s(t, (d\mu/d\theta)(t)) = x(t) \cdot (d\mu/d\theta)(t), \quad \theta\text{-a.e.}, \tag{7.6}$$

where

$$r(t, x(t)) = \begin{cases} 0 & \text{if } x(t) \in X(t), \\ +\infty & \text{if } x(t) \notin X(t). \end{cases} \tag{7.7}$$

(In view of Lemma 3, we have $x \in S$ if $r(t, x(t))$ is finite θ-almost everywhere, as implied by (7.6).) For arbitrary $x \in \mathscr{A}$ and $p \in \mathscr{B}$ (with θ depending on p as above), it would be true that

$$r(t, x(t)) + s(t, (d\mu/d\theta)(t)) \geqq x(t) \cdot (d\mu/d\theta)(t) \quad \theta\text{-a.e.} \tag{7.8}$$

Multiplying both sides of (7.5) by $(dt/d\theta)(t)$, adding this inequality to (7.8) and integrating with respect to θ, we obtain, since

$$L_0(t, x, \dot{x}) + r(t, x) = L_0(t, x, \dot{x}), \tag{7.9}$$

the inequality

$$\begin{aligned} &\int_0^1 L_0(t, x(t), \dot{x}(t))\, dt + \int_0^1 M_0(t, p(t)\dot{p}(t))\, dt + \int_0^1 s(t, d\mu(t)) \\ &\qquad \geqq \int_0^1 x(t)\, dp(t) + \int_0^1 \dot{x}(t) \cdot p(t)\, dt. \end{aligned} \tag{7.10}$$

Thus (7.10) holds for arbitrary $x \in \mathscr{A}$ and $p \in \mathscr{B}$, with equality if and only if (7.4) and (7.6) are satisfied. We note next that the transversality condition (5.6) can be expressed as

$$l(x(0), x(1)) + m(p(0), p(1)) = x(0) \cdot p(0) - x(1) \cdot p(1), \tag{7.11}$$

where for arbitrary $x \in \mathscr{A}$ and $p \in \mathscr{B}$ we would have

$$l(x(0), x(1)) + m(p(0), p(1)) \geqq x(0) \cdot p(0) - x(1) \cdot p(1). \tag{7.12}$$

(This is immediate from the definitions.) Adding (7.12) to (7.10), we get the inequality

$$F_0(x) + \bar{G}(p) \geqq 0 \tag{7.13}$$

(F_0 as in (4.21) and (4.22)) for arbitrary $x \in \mathscr{A}$ and $p \in \mathscr{B}$, with equality if and only if (7.4), (7.6) and (7.11) hold. Since the latter conditions are equivalent to the generalized optimality conditions, Lemma 8 is proved.

COROLLARY. *The functions $p \in \mathscr{B}$ in Theorem 5 are precisely the solutions to the generalized dual problem.*

It remains only to show that the solutions to the generalized dual problem can be construed as limits of minimizing (generalized) sequences in the previous dual problem. This is a corollary of the following theorem. Here, by the *weak* topology* on $\mathscr{B}$, we mean the topology induced by the linear functionals

$$p \to a \cdot p(0) + \int_0^1 y(t)\, dp(t), \quad (a, y) \in R^n \times \mathscr{C}. \tag{7.14}$$

THEOREM 6. *Let the assumptions of integrability, interiority and boundedness be satisfied. Then $\bar{G}$ is the lower semicontinuous extension of G to $\mathscr{B}$ in the weak* topology. In other words, for each $p \in \mathscr{B}$ one has*

$$\bar{G}(p) = \liminf G(p_i), \quad p_i \in \mathscr{A}, \quad p_i \to p, \tag{7.15}$$

where the limit is taken over all weak-convergent generalized sequences.*

Proof. For each $(a, y) \in R^n \times \mathscr{C}$, let $\varphi(a, y)$ denote the infimum of

$$\int_0^1 f(t, x(t) + y(t), \dot{x}(t) - A(t)[x(t) + y(t)])\, dt + l(x(0) + a, x(1)) \tag{7.16}$$

over all $x \in \mathscr{A}$ satisfying

$$x(t) + y(t) \in X(t) \quad \text{for every } t. \tag{7.17}$$

We shall demonstrate that

$$\varphi(a, y) = \sup_{p \in \mathscr{A}} \left\{ a \cdot p(0) + \int_0^1 y(t)\, dp(t) - G(p) \right\}, \tag{7.18}$$

while on the other hand, for every $p \in \mathscr{B}$,

$$\bar{G}(p) = \sup_{a \in R^n} \sup_{y \in \mathscr{C}} \left\{ a \cdot p(0) + \int_0^1 y(t)\, dp(t) - \varphi(a, y) \right\}. \tag{7.19}$$

This will imply by the fundamental theorem on conjugate convex functions (see [5]) that $\bar{G}$ is the lower semicontinuous extension of G to $\mathscr{B}$ in the weak* topology.

As a matter of fact, (7.18) is a case of Theorem 3. For the functions

$$f^y(t, x, u) = f(t, x + y(t), u - A(t)y(t)), \tag{7.20}$$

$$l^a(x(0), x(1)) = l(x(0) + a, x(1)), \tag{7.21}$$

and sets

$$X^y(t) = X(t) - y(t) \quad \text{(translation)}, \tag{7.22}$$

$\varphi(a, y)$ denotes the infimum in the control problem, where

$$\int_0^1 f^y(t, x(t), \dot{x}(t) - A(t)x(t))\, dt + l^a(x(0), x(1)) \tag{7.23}$$

is minimized over all $x \in \mathscr{A}$ satisfying

$$x(t) \in X^y(t) \quad \text{for every } t. \tag{7.24}$$

In the corresponding dual control problem, we have

$$g^y(t, p, w) = g(t, p, w) - (w - A^*(t)p) \cdot y(t), \tag{7.25}$$

$$m^a(p(0), p(1)) = m(p(0), p(1)) - a \cdot p(0). \tag{7.26}$$

Thus the dual problem consists of minimizing

$$G(p) - a \cdot p(0) - \int_0^1 y(t) \cdot \dot{p}(t)\, dt \tag{7.27}$$

over all $p \in \mathscr{A}$, so that (7.18) is just equation (4.19) in Theorem 3 in the case of f^y, l^a and $X^y(t)$. The interiority assumption in Theorem 3 is satisfied by $X^y(t)$, since it is satisfied by $X(t)$ and the function y is continuous. The boundedness assumption is likewise satisfied: the recession functions of $f^y(t, \cdot, \cdot)$ and l^a and the recession cones of the sets $X^y(t)$ are actually identical with those of $f(t, \cdot, \cdot)$, l and $X(t)$. As for the integrability assumption, the question is whether the function

$$\begin{aligned} h^y(t, x, p) &= \sup_u \{p \cdot u - f^y(t, x, u)\} \\ &= h(t, x + y(t), p) + p \cdot A(t)y(t) \end{aligned} \tag{7.28}$$

is summable in t for each fixed $x \in R^n$ and $p \in R^n$. The last term in (7.28) is summable in t, because $A(t)$ is summable in t and $y(t)$ is continuous in t. According to our integrability assumption on h, $h(t, x, p)$ is finite and summable in t for each x and p. Since the function $x \to -h(t, x, p)$ is convex, this implies [14, Cor. 2A] that $h(t, x(t), p(t))$ is summable in t for all bounded, measurable functions x and p. Hence, in particular, $h(t, x + y(t), p)$ is summable in t for $x \in R^n$, $y \in \mathscr{C}$ and $p \in R^n$, and it follows that $h^y(t, x, p)$ is summable in t. This verifies that the hypothesis of Theorem 3 is satisfied for f^y, l^a and $X^y(t)$, and equation (7.18) is thereby established.

We now turn to the proof of (7.19), which is by direct calculation employing results in [14]. The supremum in (7.19) is

$$\sup_{a \in R^n} \sup_{y \in \mathscr{C}} \sup_{x \in \mathscr{A}} \left\{ a \cdot p + \int_0^1 y(t)\, dp(t) + \int_0^1 L_0(t, x(t) + y(t), \dot{x}(t))\, dt + l(x(0) + a, x(1)) \right\} \tag{7.29}$$

(cf. proof of Theorem 3), where L_0, as before, is given by (4.20). This can also be written as

$$\begin{aligned} &\sup_{c_0 \in R^n} \sup_{z \in \mathscr{C}} \sup_{x \in \mathscr{A}} \left\{ (c_0 - x(0)) \cdot p(0) + \int_0^1 (z(t) - x(t))\, dp(t) - \int_0^1 L_0(t, z(t), \dot{x}(t))\, dt - l(c_0, x(1)) \right\} \\ &= \sup_{c_0 \in R^n} \sup_{z \in \mathscr{C}} \sup_{x \in \mathscr{A}} \left\{ c_0 \cdot p(0) - x(1) \cdot p(1) + \int_0^1 z(t)\, dp(t) + \int_0^1 p(t) \cdot \dot{x}(t)\, dt - \int_0^1 L_0(t, z(t), y(t))\, dt - l(c_0, c_1) \right\} \\ &= m(p(0), p(1)) + \sup_{z \in \mathscr{C}} \left\{ \int_0^1 z(t)\, dp(t) + Q(z) \right\}, \end{aligned} \tag{7.30}$$

where

$$Q(z) = \sup_{x \in \mathscr{A}} \left\{ \int_0^1 p(t) \cdot \dot{x}(t)\, dt - \int_0^1 L_0(t, z(t), \dot{x}(t))\, dt \right\}. \tag{7.31}$$

Let H_0 be the function in (5.7), so that

$$H_0(t, z(t), p(t)) = \sup_{v \in R^n} \{p(t) \cdot v - L_0(t, z(t), v)\}. \tag{7.32}$$

We claim that

$$Q(z) = \int_0^1 H_0(t, z(t), p(t))\, dt, \qquad z \in \mathscr{C}, \quad p \in \mathscr{B}. \tag{7.33}$$

The verification uses the fact, already remarked earlier in the proof, that $h(t, z(t), p(t))$ is summable in t for all bounded, measurable functions z and p (and hence for all $z \in \mathscr{C}$ and $p \in \mathscr{B}$). This implies, first of all, that if $z \in \mathscr{C}$ and

$$z(t) \in X(t) \quad \text{for every } t, \tag{7.34}$$

the function $(t, p) \to H(t, z(t), p)$ is summable in $t \in [0, 1]$ as well as finite and convex in $p \in R^n$. It follows then from [14, Cor. 2A] and (7.32) that

$$\int_0^1 H_0(t, z(t), p(t))\, dt = \sup \left\{ \int_0^1 [p(t) \cdot v(t) - L_0(t, z(t), v(t))]\, dt \right\} \tag{7.35}$$

for every bounded, measurable function p, where the supremum is taken over all *summable* functions $v : [0, 1] \to R^n$. Therefore, (7.33) is valid if (7.34) holds. On the other hand, if (7.34) does not hold, then the set of t values for which $z(t) \notin X(t)$ is of positive measure, due to our interiority assumption (Lemma 3). Then the integral in (7.33) is unambiguously $-\infty$, while the integral of L_0 in (7.31) is unambiguously $+\infty$ for every $x \in \mathscr{A}$. The latter implies that the supremum defining $Q(z)$ is $-\infty$. Thus (7.33) is valid even if (7.34) does not hold.

Summarizing to this point, we have shown that the supremum in (7.19) can be written as

$$m(p(0), p(1)) + \sup_{z \in \mathscr{C}} \left\{ \int_0^1 z(t)\, dp(t) + \int_0^1 H_0(t, z(t), p(t))\, dt \right\} \tag{7.36}$$

for arbitrary $p \in \mathscr{B}$. Now let

$$q(t, z) = -H_0(t, z, p(t)), \tag{7.37}$$

so that

$$\{z \in R^n | q(t, z) < +\infty\} = X(t). \tag{7.38}$$

The function q is convex in $z \in R^n$ (since $h(t, x, p)$ is concave in x), and the supremum in (7.36) is

$$\sup_{z \in \mathscr{C}} \left\{ \int_0^1 z(t)\, dp(t) - \int_0^1 q(t, z(t))\, dt \right\}. \tag{7.39}$$

To calculate the latter, we use [14, Thm. 5]. The hypothesis of this theorem requires q to be lower semicontinuous ($\not\equiv +\infty$) as a function of z for each t, measurable on $[0, 1] \times R^n$ with respect to the σ-field generated by products of Lebesgue sets in $[0, 1]$ and Borel sets in R^n, and

$$\int_V |q(t, z)|\, dt < +\infty \quad \text{for } z \in Z \tag{7.40}$$

whenever $V \subset [0, 1]$ and $Z \subset R^n$ are open sets such that $Z \subset X(t)$ for all $t \in T$. (In addition, the set (7.38) is required to satisfy conditions equivalent to our interiority assumption here; see [14, Lemma 2].)

Postponing for a moment the verification of these properties of q, we note that the theorem in question asserts

$$\begin{aligned} &\sup_{z \in \mathscr{C}} \left\{ \int_0^1 z(t)\, dp(t) - \int_0^1 q(t, z(t))\, dt \right\} \\ &\quad = \int_0^1 q^*(t, \dot{p}(t))\, dt + \int_0^1 \hat{q}^*(t, d\mu(t)), \end{aligned} \tag{7.41}$$

where $q^*(t, \cdot)$ is the convex function conjugate to $q(t, \cdot)$, $\hat{q}^*(t, \cdot)$ denotes the recession function of $q^*(t, \cdot)$, and $d\mu$ is the singular part of dp. In fact,

$$\hat{q}^*(t, w) = \sup_{x \in X(t)} w \cdot x = s(t, w) \tag{7.42}$$

in view of (7.38) [13, Thm. 13.3], while by definition,

$$\begin{aligned} q^*(t, w) &= \sup_{z \in R^n} \{z \cdot w - q(t, s)\} \\ &= g^*(t, p(t), w + A^*(t)p(t)) \end{aligned} \tag{7.43}$$

(see (5.7) and (4.11)). Thus (7.41), inserted in (7.36), yields the expression

$$m(p(0), p(1)) + \int_0^1 g^*(t, p(t), \dot{p}(t) + A^*(t)p(t))\, dt + \int_0^1 s(t, d\mu(t)) \tag{7.44}$$

for the supremum in (7.19), and this is $\bar{G}(p)$, as desired.

We complete the proof of Theorem 6 by verifying that q does have the properties listed above (preceding (7.40)). The lower semicontinuity of $q(t, z)$ in $z \in R^n$ follows from (7.32): since $L_0(t, \cdot, \cdot)$ is a lower semicontinuous, convex function on $R^n \times R^n$, this formula implies that $H_0(t, \cdot, \cdot)$ is a "lower closed" concave-convex function on $R^n \times R^n$ [13, p. 357]. This "closedness" of $H_0(t, \cdot, \cdot)$ entails upper semicontinuity in the concave argument, because H_0 nowhere has the value $+\infty$ [13, pp. 356–357]. To see the measurability of q in the required sense, we represent

$$q(t, z) = q_1(t, z) + q_2(t, z), \tag{7.45}$$

where

$$q_1(t, z) = -h(t, z, p(t)) - p(t) \cdot A(t)z \tag{7.46}$$

and

$$q_2(t, z) = \begin{cases} 0 & \text{if } z \in X(t), \\ +\infty & \text{if } z \notin X(t). \end{cases} \tag{7.47}$$

Recalling that $h(t, z, p(t))$ is summable in z (since p, being a function in $\mathscr{B}$, is measurable and bounded), we observe that $q_1(t, z)$ is certainly measurable in t for fixed $z \in R^n$, as well as finite and convex in z for fixed $t \in [0, 1]$. It follows that q_1 is measurable on $[0, 1] \times R^n$ [15, Prop. 1, ff]. On the other hand, the set

$$\{(t, z) | z \in X(t)\} \subset [0, 1] \times R^n \tag{7.48}$$

is measurable in the specified sense, because, under our interiority assumption, it is a countable intersection of open sets—this has already been shown in the proof of Theorem 3, following display (4.26). Thus the function q_2 is also measurable, and the measurability of $q = q_1 + q_2$ may be concluded. Finally, we remark that the summability of $h(t, z, p(t))$ in t for each $z \in R^n$ trivially implies the summability property required of q. Theorem 6 is now proved.

COROLLARY. *Let the assumptions of integrability, interiority and boundedness be satisfied. Then a function $p \in \mathscr{B}$ solves the generalized dual problem if and only if p is the weak* limit of a generalized sequence (p_i) in $\mathscr{A}$ which is a minimizing sequence*

in the earlier dual problem; that is,

$$\lim_i G(p_i) = \inf_{\mathcal{A}} G. \tag{7.49}$$

REFERENCES

[1] 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.

[2] J. DIEUDONNÉ, *Sur la separation des ensembles convexes*, Math. Ann., 163 (1966), pp. 1–3.

[3] J. E. FUNK AND E. G. GILBERT, *Some sufficient conditions for optimality in control problems with state space constraints*, this Journal, 8 (1970), pp. 498–504.

[4] L. M. GRAVES, *The Theory of Functions of a Real Variable*, McGraw-Hill, New York, 1956.

[5] J. J. MOREAU, *Fonctionelles convexes*, mimeographed lecture notes, Séminaire sur les equations aux derivées partielles, Collège de France, 1966–1967.

[6] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of problems. II: Applications*, this Journal, 5 (1967), pp. 90–137.

[7] ———, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

[8] ———, *Sufficiency conditions and a duality theory for mathematical programming in arbitrary linear spaces*, Nonlinear Programming, J. B. Rosen et al., eds., Academic Press, New York, 1970, pp. 323–348.

[9] C. OLECH, *Existence theorems for optimal problems with vector-valued cost function*, Trans. Amer. Math. Soc., 136 (1969), pp. 159–180.

[10] R. T. ROCKAFELLAR, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.

[11] ———, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math., 33 (1970), pp. 411–427.

[12] ———, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.

[13] ———, *Convex Analysis*. Princeton Univ. Press, Princeton, N.J., 1970.

[14] ———, *Integrals which are convex functionals, II*, Pacific J. Math., 39 (1971), pp. 439–469.

[15] ———, *Convex integral functionals and duality*, Contributions to Nonlinear Functional Analysis, E. Zarantonello, ed., Academic Press, New York, 1971, pp. 215–235.

[16] M. M. TSVETANOV, *Duality in problems of the calculus of variations and optimal control*, Doctoral thesis, Moscow State University, Moscow, 1970.

[17] V. L. MAKAROV, *A property of solutions of a problem of continuous linear and convex programming*, Dokl. Akad. Nauk SSSR, 176 (1967)= Soviet Math. Dokl., 8 (1967), pp. 1246–1247.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

STABILIZATION OF LINEAR SYSTEMS*

MASAO IKEDA,† HAJIME MAEDA‡ AND SHINZO KODAMA†

Abstract. This paper considers a finite-dimensional linear time-varying system and is concerned with the question: What is the relation between controllability properties of the system and various degrees of stability of the closed loop system resulting from linear feedback of the state variable?

The main results are as follows: For any initial time t_0, and any continuous and monotonically nondecreasing function $\delta(\cdot, t_0)$ such that $\delta(t_0, t_0) = 0$, the transition matrix $\hat{\Phi}(\cdot, \cdot)$ of the closed loop system can be made such that $\|\hat{\Phi}(t, t_0)\| \leqq a(t_0) \exp[-\delta(t, t_0)]$ for all $t \geqq t_0$, if and only if the system is completely controllable. Furthermore, in case of a bounded system, for any $m \geqq 0$, a bounded feedback matrix can be found such that $\|\hat{\Phi}(t_2, t_1)\| \leqq a \exp[-m(t_2 - t_1)]$ for all t_1 and $t_2 \geqq t_1$, if and only if the system is uniformly completely controllable.

Thus they can be regarded as extensions of the well-known result of Wonham [1] for a time-invariant system (i.e., the equivalence between complete controllability and the possibility of closed loop pole assignment), and also the results of Kalman [2], Johnson [3] and Anderson and Moore [4] for a time-varying system in which sufficient conditions for stabilization of the closed loop system are given.

1. Introduction. This paper considers the problem of stabilizing a finite-dimensional linear time-varying system by means of linear feedback of the state. The object is to clarify the relation between some concepts of controllability (i.e., complete controllability and uniform complete controllability) of open loop systems and various degrees of stability of the closed loop system which can be obtained by the state variable feedback. For the time-invariant case, it is well known (Wonham [1]) that controllability of the open loop system is equivalent to the possibility of assigning an arbitrary set of the closed-loop-system poles. For the time-varying case, Kalman [2] has shown that the closed loop system can be made uniformly asymptotically stable if the open loop system is uniformly completely controllable. Under the same condition, Johnson [3] and Anderson and Moore [4] have pointed out the possibility of making the zero-input response of the closed loop system decay faster than $\exp[-mt]$ for any specified real m. Wolovich [5] has considered a class of systems for which a Lyapunov transformation can be found which leads to a time-invariant system, and has shown from a viewpoint of pole assignment that a slightly more restrictive type of controllability than uniform controllability is a sufficient condition for uniform asymptotic stabilization.

This paper is also concerned with a finite-dimensional linear time-varying system and the object is to investigate equivalence between the open loop system controllability and the closed loop system stability. The results can be roughly stated as follows: The closed loop zero-input response starting at t_0 can be given any decay characteristic of the form $a(t_0) \exp[-\delta(t, t_0)]$, where $\delta(t, t_0)$, $t \geqq t_0$, is an arbitrary continuous and nondecreasing function of t, if and only if the open loop system is completely controllable. If the open loop system is bounded, then

* Received by the editors March 30, 1971, and in final revised form December 20, 1971.

† Department of Communication Engineering, Faculty of Engineering, Osaka University, Suita, Osaka, Japan.

‡ Department of Mechanical Engineering, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka, Japan.

the closed loop zero-input response can be made to decay faster than $\exp[-mt]$ for any specified real m by a bounded feedback gain if and only if the open loop system is uniformly completely controllable. Thus results of this paper constitute logical extensions of Wonham [1] and supplement the results of Kalman [2], Johnson [3] and Anderson and Moore [4].

2. Notations and definitions. Throughout the paper, all vectors and matrices are assumed to have real elements. For a matrix A, A' is the transpose, $\lambda_{\max}(A)$ ($\lambda_{\min}(A)$) is the maximum (minimum) eigenvalue of A, $\operatorname{tr}(A)$ is the trace of A, and $\|A\|$ denotes the norm of A compatible with the Euclidean norm $\|x\| = (x'x)^{1/2}$ of a vector x; that is, $\|A\| = \sup_{\|x\|=1} \|Ax\| = (\lambda_{\max}(A'A))^{1/2}$.

For symmetric matrices Q and R, $Q > 0$ ($Q \geqq 0$) means Q is positive (nonnegative) definite, and $Q > R$ ($Q \geqq R$) means $Q - R > 0$ ($Q - R \geqq 0$).

Consider the matrix Riccati equation

$$\dot{P}(t) + F'(t)P(t) + P(t)F(t) - P(t)G(t)G'(t)P(t) = -D'(t)D(t), \quad t \geqq t_r,$$

where matrices $F(t)$, $G(t)$ and $D(t)$ are measurable and bounded on every finite subinterval of $[t_r, \infty)$. Corresponding to a terminal condition $P(T) = P_T \geqq 0$ at $t = T > t_r$, there is a unique, continuous, symmetric and nonnegative definite solution $P(t; P_T, T)$ on $[t_r, T]$ [6]. If $\lim_{T\to\infty} P(t; 0, T)$ exists for all $t \geqq t_r$, this is also a solution [2], which will be denoted by $P_0(t)$ in the sequel.

We assume that the open loop system has a representation

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \tag{1}$$

where $x(t)$ is the state n-vector, $u(t)$ is the input p-vector, matrices $A(t)$ and $B(t)$ are respectively $n \times n$ and $n \times p$. In addition, $u(t)$, $A(t)$ and $B(t)$ are assumed to be measurable and bounded on every finite subinterval of time. The representation (1) will be denoted by $R(A(t), B(t))$ in the sequel. Consider linear state variable feedback of the form

$$u(t) = -K(t)x(t) + v(t), \tag{2}$$

where the feedback gain $K(t)$ is $p \times n$, measurable and bounded on every finite subinterval of the interval on which it is defined, and $v(t)$ is the new input. Then the closed loop system is represented by

$$\dot{x}(t) = \hat{A}(t)x(t) + B(t)v(t), \quad \hat{A}(t) = A(t) - B(t)K(t) \tag{3}$$

which will be denoted as $R(\hat{A}(t), B(t))$ in the following. The transition matrices associated with $\dot{x}(t) = A(t)x(t)$ and $\dot{x}(t) = \hat{A}(t)x(t)$ are denoted by $\Phi(\cdot, \cdot)$ and $\hat{\Phi}(\cdot, \cdot)$ respectively. In view of the assumptions on $A(t)$, $B(t)$ and $K(t)$, unique, continuous, nonsingular $\Phi(\cdot, \cdot)$ and $\hat{\Phi}(\cdot, \cdot)$ exist [7].

The following definitions of controllability are due to Kalman [2].

DEFINITION 1. A state x of $R(A(t), B(t))$ is said to be *controllable at time* t_i, if there exists an input defined over some finite closed interval $[t_i, t_f]$ which transfers the state from x at t_i to the origin 0 at t_f. If every state of $R(A(t), B(t))$ is controllable at t_i, it is said to be *completely controllable at* t_i. If $R(A(t), B(t))$ is completely controllable at every time, it is said to be *completely controllable.*

DEFINITION 2. $R(A(t), B(t))$ is said to be *uniformly completely controllable*, if there is a fixed number $\sigma > 0$ such that

$$0 < \alpha_1(\sigma)I \leqq W(t, t+\sigma) \leqq \alpha_2(\alpha)I, \tag{4a}$$

$$0 < \alpha_3(\sigma)I \leqq \Phi(t+\sigma, t)W(t, t+\sigma)\Phi'(t+\sigma, t) \leqq \alpha_4(\sigma)I \tag{4b}$$

hold for all t, where $W(t, t+\sigma)$ is a symmetric matrix defined by

$$W(t, t+\sigma) = \int_t^{t+\sigma} \Phi(t, s)B(s)B'(s)\Phi'(t, s)\, ds. \tag{5}$$

Remark 1. It is shown in [2] that if (4a) and (4b) hold for some σ, then there is a function $\alpha_5(\cdot)$ such that

$$\|\Phi(t, s)\| \leqq \alpha_5(|t - s|) \quad \text{for all } t \text{ and } s.$$

This implies that

$$0 < \beta_1(|s - t|)I \leqq \Phi'(s, t)\Phi(s, t) \leqq \beta_2(|s - t|)I \quad \text{for all } s \text{ and } t,$$

where $\beta_1(\cdot) = \alpha_5^{-2}(\cdot)$ and $\beta_2(\cdot) = \alpha_5^2(\cdot)$.

3. Concepts of stabilizability. Let $\delta(t, t_0)$ denote a continuous and monotonically nondecreasing function which is defined for all $t \geqq t_0$ and satisfies $\delta(t_0, t_0) = 0$. $\delta(t, t_0)$ will serve as a measure of decay of the zero-input response of the closed loop system.

DEFINITION 3. $R(A(t), B(t))$ is said to be *completely stabilizable* if, for any initial time t_0 and any $\delta(t, t_0)$, there exist a feedback gain $K(t)$ defined for $t \geqq t_0$ and a positive number $a(t_0)$ such that

$$\|\hat{\Phi}(t, t_0)\| \leqq a(t_0) \exp[-\delta(t, t_0)] \quad \text{for all } t \geqq t_0. \tag{6}$$

DEFINITION 4. $R(A(t), B(t))$ is said to be *uniformly completely stabilizable* if, for any nonnegative m, there exist a feedback gain $K(t)$ defined for all t and a positive number a such that

$$\|\hat{\Phi}(t_2, t_1)\| \leqq a \exp[-m(t_2 - t_1)] \quad \text{for all } t_1 \text{ and } t_2 \geqq t_1. \tag{7}$$

Remark 2. If $R(A(t), B(t))$ is completely stabilizable, the feedback gain $K(t)$ generally depends on the initial time t_0 and the specified function $\delta(t, t_0)$. In case of a uniformly completely stabilizable system, $K(t)$ depends only on m.

Remark 3. Without loss of generality, we may assume that $\delta(t, t_0)$ is continuously differentiable (with respect to the first variable) for the following reason. For each $\delta(t, t_0)$, there is always a monotonically nondecreasing and continuously differentiable function $\delta_d(t, t_0)$ such that $\delta_d(t_0, t_0) = 0$ and $\delta_d(t, t_0) \geqq \delta(t, t_0)$ for all $t \geqq t_0$. Thus given any $\delta(t, t_0)$, we replace it by $\delta_d(t, t_0)$ and find $K(t)$. This feedback gain achieves the required stability specified by $\delta(t, t_0)$. Conversely, if a feedback gain can be found for any $\delta(t, t_0)$, naturally it should be found for any continuously differentiable $\delta(t, t_0)$.

The relation between the concepts of stabilizability are illustrated by the following two examples. That is, complete stabilizability does not imply uniform complete stabilizability and the converse is not true either. We can expect this

fact because of the difference between the two definitions. For example, complete stabilizability does not guarantee that $a(t_0)$ and $K(t)$ become independent of the initial time t_0 even when $\delta(t, t_0) = m(t - t_0)$; uniform complete stabilizability requires that both $a(t_0)$ and $K(t)$ be independent of t_0. Conversely, uniform complete stabilizability only considers stabilization with respect to the class of exponential functions, while complete stabilizability considers a much broader class of functions.

Example 1. Consider a scalar system

$$\dot{x}(t) = x(t) + 1(t)u(t), \tag{8}$$

where $1(t)$ is the unit step function. For any t_0 and any $\delta(t, t_0)$, define

$$K(t) = \dot{\delta}(t, t_0) + 1, \qquad t \geqq t_0.$$

Then the solution of (8) satisfies

$$|x(t)| \leqq a(t_0)|x(t_0)| \exp[-\delta(t, t_0)] \quad \text{for all } t \geqq t_0,$$

where

$$a(t_0) = \begin{cases} \exp[-t_0 + \delta(0, t_0)], & t_0 < 0, \\ 1, & t_0 \geqq 0. \end{cases}$$

This shows that (8) is completely stabilizable. Note however that when $t_1 < 0$, $x(0) = x(t_1)\exp[-t_1]$ for any feedback gain. Thus necessarily $a \geqq \exp[-t_1]$. This implies that (8) is not uniformly completely stabilizable.

Example 2. Consider

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -2|t| \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t). \tag{9}$$

For any $m \geqq 0$, define

$$K(t) = [m + 1 \quad 0].$$

Then

$$x_1(t_2) = x_1(t_1)\exp[-m(t_2 - t_1)],$$
$$x_2(t_2) = x_2(t_1)\exp[-(t_2^3/|t_2|) + (t_1^3/|t_1|)]$$

for all t_1 and $t_2 \geqq t_1$. Since, for any $m \geqq 0$,

$$-\frac{t_2^3}{|t_2|} + \frac{t_1^3}{|t_1|} \leqq \frac{m^2}{2} - m(t_2 - t_1) \quad \text{for all } t_1 \text{ and } t_2 \geqq t_1,$$

the solution of (9) satisfies

$$\|x(t_2)\| \leqq a\|x(t_1)\| \exp[-m(t_2 - t_1)] \quad \text{for all } t_1 \text{ and } t_2 \geqq t_1,$$

where $a = \exp[m^2/2]$. This shows that (9) is uniformly completely stabilizable. It can be shown (by Theorem 1) that it is not completely stabilizable.

4. Main results. Results of this paper are summarized in three theorems. Theorem 1 can be regarded as an extension of Wonham [1] to a general time-varying system while Theorem 3 is an extension to a bounded case. Theorem 2

indicates that the boundedness of $A(t)$ and $B(t)$ can be removed from the sufficiency part of Theorem 3 if $K(t)$ is not required to be bounded.

THEOREM 1. *$R(A(t), B(t))$ is completely stabilizable if and only if it is completely controllable.*

THEOREM 2. *If $R(A(t), B(t))$ is uniformly completely controllable, then it is uniformly completely stabilizable.*

THEOREM 3. *When $R(A(t), B(t))$ is a bounded representation (i.e., $\|A(t)\|$ and $\|B(t)\|$ are bounded), it is uniformly completely stabilizable by a bounded feedback gain if and only if it is uniformly completely controllable.*[1]

Remark. 4. The converse of Theorem 2 is not valid as shown by the following counterexample.

Example 3. Consider a scalar system

$$\dot{x}(t) = x(t) + \exp[2t]u(t). \tag{10}$$

For any $m \geqq 0$, define

$$K(t) = (m + 1)\exp[-2t].$$

Then

$$x(t_2) = x(t_1)\exp[-m(t_2 - t_1)] \quad \text{for all } t_1 \text{ and } t_2 \geqq t_1.$$

This shows that (10) is uniformly completely stabilizable, but it is not uniformly completely controllable since the controllability matrix defined by (5) is

$$W(t, t + \sigma) = \tfrac{1}{2}(\exp[2\sigma] - 1)\exp[4t] \quad \text{for each } \sigma > 0.$$

Remark 5. If $R(A(t), B(t))$ is not bounded, Theorem 3 is not valid. This is shown by Example 2, that is, (9) is not (uniformly) completely controllable, but it is uniformly completely stabilizable by a bounded feedback gain.

5. Proofs of theorems.

5.1. Preliminaries for the proof of Theorem 1. In the proof of necessity, the structural decomposition of $R(A(t), B(t))$ obtained by Kalman, Ho and Narendra [9] plays an important role, for which we give an algebraic proof in the following. In this way, the relation between $R(A(t), B(t))$ and the matrices characterizing the decomposed system is clearly established.

LEMMA 1. *If $R(A(t), B(t))$ is not completely controllable, then there is a time t_u such that for all $t \geqq t_u$, the state $x(t)$ can be represented as an orthogonal direct sum of n-vectors $x_v(t)$ and $x_w(t)$ which satisfy equations of the form*

$$\begin{aligned} \dot{x}_v(t) &= A_{vv}(t)x_v(t) + A_{vw}(t)x_w(t) + B(t)u(t), \\ \dot{x}_w(t) &= A_{ww}(t)x_w(t) \end{aligned} \tag{11}$$

almost everywhere on $[t_u, \infty)$, where $A_{vv}(t)$, $A_{vw}(t)$ and $A_{ww}(t)$ are measurable and bounded on every finite subinterval of $[t_u, \infty)$.

Proof. Assume that $R(A(t), B(t))$ is not completely controllable. Then there is a time t_u when the dimension of the controllable subspace $C(t_u)$ is less than the dimension of the state space X. For all t, we define an orthogonal direct-sum decomposition of X by

$$X = V(t) \oplus W(t),$$

[1] Anderson and Moore give the sufficient part of Theorem 3 for a time-invariant system [8] and indicate that it is applicable for a bounded time-varying system [4].

where

$$V(t) = \Phi(t, t_u)C(t_u) \quad \text{and} \quad W(t) = \Phi'(t_u, t)U(t_u),$$

$U(t_u)$ being the orthogonal complement of $C(t_u)$ (hence $U(t_u)$ is the uncontrollable subspace at t_u). Let

$$x(t) = x_v(t) + x_w(t),$$

where $x_v(t) \in V(t)$ and $x_w(t) \in W(t)$. Clearly,

$$x_v'(t)x_w(t) = 0 \quad \text{for all } t.$$

Define vectors $u_1(t), \cdots, u_n(t)$ by

$$u_i(t) = \begin{cases} \Phi(t, t_u)v_i, & i = 1, \cdots, r, \\ \Phi'(t_u, t)w_{i-r}, & i = r+1, \cdots, n, \end{cases}$$

where the sets $\{v_1, \cdots, v_r\}$ and $\{w_1, \cdots, w_{n-r}\}$ are arbitrary bases of $C(t_u)$ and $U(t_u)$ respectively. Let

$$T(t) = \begin{bmatrix} u_1'(t) \\ \vdots \\ u_n'(t) \end{bmatrix}.$$

$T(t)$ is continuous and nonsingular at every t, and $\dot{T}(t)$ exists almost everywhere and is measurable and bounded on every finite subinterval of time since

$$\dot{u}_i(t) = \begin{cases} A(t)u_i(t), & i = 1, \cdots, r, \\ -A'(t)u_i(t), & i = r+1, \cdots, n, \end{cases}$$

whenever $\dot{u}_i(t)$ exists. From the definition of $u_i(t)$, the set $\{u_1(t), \cdots, u_r(t)\}$ is a basis of $V(t)$ and the set $\{u_{r+1}(t), \cdots, u_n(t)\}$ is a basis of $W(t)$. Let

$$T(t)x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix},$$

where

$$T(t)x_v(t) = \begin{bmatrix} x_1(t) \\ 0 \end{bmatrix} \begin{matrix} \}r \\ \}n-r \end{matrix} \quad \text{and} \quad T(t)x_w(t) = \begin{bmatrix} 0 \\ x_2(t) \end{bmatrix} \begin{matrix} \}r \\ \}n-r \end{matrix}.$$

Then $x_1(t)$ and $x_2(t)$ satisfy

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \bar{A}(t) \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \bar{B}(t)u(t)$$

almost everywhere, where

$$\bar{A}(t) = (\dot{T}(t) + T(t)A(t))T^{-1}(t) \quad \text{and} \quad \bar{B}(t) = T(t)B(t).$$

Here it is evident that $\bar{A}(t)$ and $\bar{B}(t)$ are measurable and bounded on every finite subinterval of time. The direct calculation shows that $\bar{A}(t)$ has the form

$$\bar{A}(t) = \begin{bmatrix} \overbrace{A_{11}(t)}^{r} & \overbrace{A_{12}(t)}^{n-r} \\ 0 & 0 \end{bmatrix} \begin{matrix} \}r \\ \}n-r \end{matrix} \quad \text{almost everywhere.}$$

The last $n - r$ row vectors of $\bar{B}(t)$ are $w_i'\Phi(t_u, t)B(t) = 0$ $(i = 1, \cdots, n - r)$ for almost all $t \geqq t_u$, because $w_i \in U(t_u)$ and $U(t_u)$ is the null space of $W(t_u, t)$ for any $t > t_u$ [9]. Then

$$\bar{B}(t) = \begin{bmatrix} \overbrace{B_1(t)}^{p} \\ 0 \end{bmatrix} \begin{matrix} \}r \\ \}n - r \end{matrix} \quad \text{for almost all } t \geqq t_u.$$

The form

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} A_{11}(t) & A_{12}(t) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} B_1(t) \\ 0 \end{bmatrix} u(t), \quad t \geqq t_u,$$

together with

$$x_v(t) = T^{-1}(t) \begin{bmatrix} x_1(t) \\ 0 \end{bmatrix} \quad \text{and} \quad x_w(t) = T^{-1}(t) \begin{bmatrix} 0 \\ x_2(t) \end{bmatrix}$$

implies that $x_v(t)$ and $x_w(t)$ satisfy the following equations almost everywhere on $[t_u, \infty)$:

$$\dot{x}_v(t) = A_{vv}(t)x_v(t) + A_{vw}(t)x_w(t) + B_v(t)u(t),$$

$$\dot{x}_w(t) = A_{ww}(t)x_w(t),$$

where

$$A_{vv}(t) = -T^{-1}(t)\dot{T}(t) + T^{-1}(t) \begin{bmatrix} A_{11}(t) & 0 \\ 0 & 0 \end{bmatrix} T(t),$$

$$A_{vw}(t) = T^{-1}(t) \begin{bmatrix} 0 & A_{12}(t) \\ 0 & 0 \end{bmatrix} T(t),$$

$$A_{ww}(t) = -T^{-1}(t)\dot{T}(t),$$

$$B_v(t) = T^{-1}(t) \begin{bmatrix} B_1(t) \\ 0 \end{bmatrix} = T^{-1}(t)\bar{B}(t) = B(t).$$

Here we find $A_{vv}(t)$, $A_{vw}(t)$ and $A_{ww}(t)$ measurable and bounded on every finite subinterval of $[t_u, \infty)$.

LEMMA 2. *If $R(A(t), B(t))$ is completely controllable, then $R(\tilde{A}(t), B(t))$ defined for $t \geqq t_0$ by $\tilde{A}(t) = A(t) + \delta(t, t_0)I$ is completely controllable at every $t \geqq t_0$.*

Proof. Let $y(t) = x(t) \exp [\delta(t, t_0)]$ for $t \geqq t_0$. Then $R(A(t), B(t))$ is transformed to

$$\dot{y}(t) = \tilde{A}(t)y(t) + B(t) \exp [\delta(t, t_0)]u(t), \quad t \geqq t_0.$$

Since complete controllability is preserved under linear nonsingular transformations [10], the above system is completely controllable at every $t \geqq t_0$. Thus $R(\tilde{A}(t), B(t))$ is completely controllable at every $t \geqq t_0$ with respect to the input $u(t) \exp [\delta(t, t_0)]$.

5.2. Proof of Theorem 1.

Necessity. Assume that $R(A(t), B(t))$ is completely stabilizable. If it is not completely controllable, then, from Lemma 1, there is a time t_u such that for all

$t \geqq t_u$, the state $x(t)$ of $\dot{x}(t) = \hat{A}(t)x(t)$ (i.e., $R(\hat{A}(t), 0)$) can be represented as an orthogonal direct sum of $x_v(t)$ and $x_w(t)$ which satisfy

$$\begin{aligned} \dot{x}_v(t) &= [A_{vv}(t) - B(t)K(t)]x_v(t) + [A_{vw}(t) - B(t)K(t)]x_w(t), \\ \dot{x}_w(t) &= A_{ww}(t)x_w(t) \end{aligned} \tag{12}$$

almost everywhere on $[t_u, \infty)$. Here, since $x_w(t)$ is orthogonal to $x_v(t)$,

$$\|x_w(t)\| \leqq \|x(t)\| \quad \text{for all } t \geqq t_u, \tag{13a}$$

and by Wazewski's theorem [11], along any solution of (12),

$$\|x_w(t_1)\| \exp\left[\frac{1}{2}\int_{t_1}^{t_2} \lambda_{\min}(A_{ww}(s) + A'_{ww}(s))\, ds\right] \leqq \|x_w(t_2)\| \tag{13b}$$

for all $t_1 \geqq t_u$ and $t_2 \geqq t_1$. Since $A_{ww}(t)$ is bounded on every finite subinterval of $[t_u, \infty)$, this inequality implies that $x_w(t)$ does not settle to the origin 0 in finite time. Let $t_0 = t_u$ and let

$$\delta(t, t_0) = t - t_u + \frac{1}{2}\int_{t_u}^{t} |\lambda_{\min}(A_{ww}(s) + A'_{ww}(s))|\, ds \quad \text{for } t \geqq t_0 = t_u. \tag{13c}$$

By assumption, there exist $K(t)$ and $a(t_0)$ such that

$$\|x(t)\| \leqq a(t_0)\|x(t_0)\| \exp[-\delta(t, t_0)] \quad \text{for all } t \geqq t_0 \tag{13d}$$

along any solution of (12). Thus (13a)–(13d) imply that for any initial state of (12),

$$\begin{aligned} &\|x_w(t_0)\| \exp\left[\frac{1}{2}\int_{t_u}^{t} \lambda_{\min}(A_{ww}(s) + A'_{ww}(s))\, ds\right] \\ &\qquad \leqq a(t_0)\|x(t_0)\| \exp\left[-(t - t_u) - \frac{1}{2}\int_{t_u}^{t} |\lambda_{\min}(A_{ww}(s) + A'_{ww}(s))|\, ds\right] \end{aligned}$$

holds for all $t \geqq t_0 = t_u$. Since $x_w(t_0)$ can be a nonzero vector, this inequality is a contradiction. Therefore, if $R(A(t), B(t))$ is completely stabilizable, then it is completely controllable.

Sufficiency. Assume that $R(A(t), B(t))$ is completely controllable. In view of Remark 3, it suffices to construct a feedback gain $K(t)$ for a continuously differentiable $\delta(t, t_0)$. Let $y(t) = x(t) \exp[\delta(t, t_0)]$ for $t \geqq t_0$. Then $R(\hat{A}(t), 0)$ is transformed to

$$\dot{y}(t) = [\tilde{A}(t) - B(t)K(t)]y(t), \quad t \geqq t_0, \tag{14}$$

where $\tilde{A}(t) = A(t) + \dot{\delta}(t, t_0)I$. Note that $x(t)$ of $R(\hat{A}(t), 0)$ satisfies the decay characteristic specified by $\delta(t, t_0)$ if and only if $y(t)$ of (14) is bounded for any bounded $y(t_0)$.

Define $K(t)$ for any t_0 and any $\delta(t, t_0)$ by

$$K(t) = \tfrac{1}{2}B'(t)[P_0(t) - I], \quad t \geqq t_0, \tag{15}$$

where the $n \times n$ matrix $P_0(t)$ is the solution of the matrix Riccati equation

$$\dot{P}(t) + \tilde{A}'(t)P(t) + P(t)\tilde{A}(t) - P(t)B(t)B'(t)P(t) = -H(t), \quad t \geqq t_0,$$

where $H(t)$ is any $n \times n$ matrix function which is measurable, bounded on every

finite subinterval of $[t_0, \infty)$, symmetric and nonnegative definite, and satisfies

$$H(t) \geqq \tilde{A}'(t) + \tilde{A}(t) + B(t)B'(t)$$

almost everywhere on $[t_0, \infty)$. The existence of such solution $P_0(t)$ is guaranteed by the assumption and Lemma 2 [2]. Since $P_0(t)$ is continuous at every t on $[t_0, \infty)$, $K(t)$ defined by (15) is measurable and bounded on every finite subinterval of $[t_0, \infty)$.

Now we shall show that this $K(t)$ is a desired feedback gain. Consider a scalar function $V(y, t)$ defined by

$$V(y, t) = y'P_0(t)y + y'y, \quad t \geqq t_0.$$

Since $P_0(t)$ is nonnegative definite, $V(y, t)$ is bounded from below by a nondecreasing function of $\|y\|$ independent of t, and $V(y, t) \to \infty$ as $\|y\| \to \infty$. The time derivative of $V(y, t)$ along any solution of (14) is nonpositive almost everywhere on $[t_0, \infty)$. Therefore, by means of Lyapunov's second method, any solution of (14) is bounded and, more precisely,

$$\|y(t)\| \leqq a(t_0)\|y(t_0)\| \quad \text{for all } t \geqq t_0,$$

where $a(t_0) = [1 + \lambda_{\max}(P_0(t_0))]^{1/2}$. Thus any solution $x(t)$ of $R(\hat{A}(t), 0)$, where $K(t)$ is given by (15), satisfies

$$\|x(t)\| \leqq a(t_0)\|x(t_0)\| \exp[-\delta(t, t_0)] \quad \text{for all } t \geqq t_0.$$

5.3. Preliminaries for the proof of Theorem 2.

LEMMA 3. *If $R(A(t), B(t))$ is completely controllable, then the solution*[2] *$Q_0(t)$ of the matrix Riccati equation*

$$\dot{Q}(t) + A'(t)Q(t) + Q(t)A(t) - Q(t)B(t)B'(t)Q(t) = -I \tag{16}$$

exists and satisfies

$$\begin{aligned} &\left[Y^{-1}(t, t_d(t)) + \operatorname{tr}(W(t, t_d(t)))\left(1 + \frac{\operatorname{tr}(Y(t, t_d(t)))}{\lambda_{\min}(Y(t, t_d(t)))}\right)^2 I \right]^{-1} \\ &\quad \leqq Q_0(t) \\ &\quad \leqq W^{-1}(t, t_c(t)) + \operatorname{tr}(Y(t, t_c(t)))\left(1 + \frac{\operatorname{tr}(W(t, t_c(t)))}{\lambda_{\min}(W(t, t_c(t)))}\right)^2 I \end{aligned} \tag{17}$$

for all t, where

$$Y(t, t_j(t)) = \int_t^{t_j(t)} \Phi'(s, t)\Phi(s, t)\, ds, \qquad j = d, c.$$

Here $W(t, t_j(t)), j = d, c$, is the controllability matrix of $R(A(t), B(t))$ defined by (5), $t_c(t)$ is the time for which $W(t, t_c(t))$ is positive definite and $t_d(t)$ is any time later than t.

Proof. The existence of $Q_0(t)$ is shown in [2], and the upper bound is essentially established in [2]. (See the proof of Stability Theorem in [2].) The lower bound is

[2] $Q_0(t)$ is defined in a manner analogous to that of $P_0(t)$ in § 2; i.e., $\lim_{T\to\infty} Q(t, 0, T)$, where $Q(t, 0, T)$ is the solution of the matrix Ricatti equation satisfying the terminal condition $Q(T) = 0$.

shown as follows. Consider the identity [2]

$$x'Q_0(t)x = \inf_u \int_t^\infty (\|x(s)\|^2 + \|u(s)\|^2)\, ds,$$

where $x(s)$, $s \geqq t$, is the solution of $R(A(t), B(t))$ starting from x at t with the input $u(\cdot)$ on $[t, s]$. This implies that $Q_0(t)$ is positive definite and hence the inverse of $Q_0(t)$ exists everywhere. Let $Q_0^{-1}(t) = R(t)$. Then $R(t)$ satisfies

$$\dot{R}(t) - A(t)R(t) - R(t)A'(t) - R(t)R(t) = -B(t)B'(t)$$

almost everywhere, which implies

$$x'R(t)x = \inf_u \left[x'(t_d(t))R(t_d(t))x(t_d(t)) + \int_t^{t_d(t)} (x'(s)B(s)B'(s)x(s) + \|u(s)\|^2)\, ds \right]$$

for all t and $t_d(t) > t$, where $x(s), t \leqq s \leqq t_d(t)$, is the solution of $R(-A'(t), I)$ starting from x at t with the input $u(\cdot)$ on $[t, s]$. The same consideration used to obtain the upper bound on $Q_0(t)$ (i.e., [2]) implies

$$R(t) \leqq W^{*-1}(t, t_d(t)) + \operatorname{tr}(Y^*(t, t_d(t)))\left(1 + \frac{\operatorname{tr}(W^*(t, t_d(t)))}{\lambda_{\min}(W^*(t, t_d(t)))}\right)^2 I$$

for all t and $t_d(t) > t$, where, if the transition matrix of $\dot{x}(t) = -A'(t)x(t)$ is denoted by $\Phi^*(\cdot, \cdot)$,

$$W^*(t, t_d(t)) = \int_t^{t_d(t)} \Phi^*(t, s)\Phi^{*\prime}(t, s)\, ds,$$

$$Y^*(t, t_d(t)) = \int_t^{t_d(t)} \Phi^{*\prime}(s, t)B(s)B'(s)\Phi^*(s, t)\, ds.$$

Note that $\Phi^*(t, s) = \Phi'(s, t)$. Then $W^*(t, t_d(t)) = Y(t, t_d(t))$ and $Y^*(t, t_d(t)) = W(t, t_d(t))$, and therefore

$$R(t) \leqq Y^{-1}(t, t_d(t)) + \operatorname{tr}(W(t, t_d(t)))\left(1 + \frac{\operatorname{tr}(Y(t, t_d(t)))}{\lambda_{\min}(Y(t, t_d(t)))}\right)^2 I$$

for all t and $t_d(t) > t$. This is equivalent to the lower bound.

LEMMA 4. *If $R(A(t), B(t))$ is uniformly completely controllable, then there exist two functions $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ such that*

$$0 < \gamma_1(t_2 - t_1)I \leqq \int_{t_1}^{t_2} \Phi'(s, t_1)\Phi(s, t_1)\, ds \leqq \gamma_2(t_2 - t_1)I$$

for all t_1 and $t_2 > t_1$.

Proof. In view of Remark 1, let $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ be defined by

$$\gamma_1(t_2 - t_1) = \int_{t_1}^{t_2} \beta_1(|s - t_1|)\, ds = \int_0^{t_2 - t_1} \beta_1(s)\, ds,$$

$$\gamma_2(t_2 - t_1) = \int_{t_1}^{t_2} \beta_2(|s - t_1|)\, ds = \int_0^{t_2 - t_1} \beta_2(s)\, ds.$$

Then the inequalities follow.

LEMMA 5. *If $R(A(t), B(t))$ is uniformly completely controllable, then the solution $Q_0(t)$ of (16) is bounded from above and below by positive definite matrices independent of t.*[3]

Proof. Let $t_c(t) = t_d(t) = t + \sigma$ in (17), where $\sigma > 0$ is the fixed number in Definition 2. Then, from the assumption and Lemma 4,

$$0 < \left[\frac{1}{\gamma_1(\sigma)} + n\alpha_2(\sigma)\left(1 + \frac{n\gamma_2(\sigma)}{\gamma_1(\sigma)}\right)^2\right]^{-1} I$$
$$\leqq Q_0(t) \leqq \left[\frac{1}{\alpha_1(\sigma)} + n\gamma_2(\sigma)\left(1 + \frac{n\alpha_2(\sigma)}{\alpha_1(\sigma)}\right)^2\right] I \quad \text{for all } t.$$

LEMMA 6. *If $R(A(t), B(t))$ is uniformly completely controllable, then $R(A(t) + mI, B(t))$, where m is any nonnegative number, is also uniformly completely controllable.*[4]

Proof. Note that the transition matrix $\tilde{\Phi}(t, s)$ of $\dot{x}(t) = [A(t) + mI]x(t)$ is represented by $\Phi(t, s) \exp [m(t - s)]$. The controllability matrix

$$\tilde{W}(t, t + \sigma) = \int_t^{t+\sigma} \tilde{\Phi}(t, s)B(s)B'(s)\tilde{\Phi}'(t, s)\, ds$$

of $R(A(t) + mI, B(t))$ satisfies

$$W(t, t + \sigma) \exp [-2m\sigma] \leqq \tilde{W}(t, t + \sigma) \leqq W(t, t + \sigma) \quad \text{for all } t,$$

since

$$\Phi(t, s)B(s)B'(s)\Phi'(t, s) \exp [-2m\sigma] \leqq \tilde{\Phi}(t, s)B(s)B'(s)\tilde{\Phi}'(t, s)$$
$$\leqq \Phi(t, s)B(s)B'(s)\Phi'(t, s),$$

$t \leqq s \leqq t + \sigma$. Similarly,

$$\Phi(t + \sigma, t)W(t, t + \sigma)\Phi'(t + \sigma) \leqq \tilde{\Phi}(t + \sigma, t)\tilde{W}(t, t + \sigma)\tilde{\Phi}'(t + \sigma, t)$$
$$\leqq \Phi(t + \sigma, t)W(t, t + \sigma)\Phi'(t + \sigma, t) \exp [2m\sigma]$$

for all t, since

$$\Phi(t + \sigma, s)B(s)B'(s)\Phi'(t + \sigma, s) \leqq \tilde{\Phi}(t + \sigma, s)B(s)B'(s)\tilde{\Phi}'(t + \sigma, s)$$
$$\leqq \Phi(t + \sigma, s)B(s)B'(s)\Phi'(t + \sigma, s) \exp [2m\sigma],$$

$t \leqq s \leqq t + \sigma$. Therefore, from the assumption, there exists a fixed number $\sigma > 0$ such that

$$0 < \alpha_1(\sigma) \exp [-2m\sigma]I \leqq \tilde{W}(t, t + \sigma) \leqq \alpha_2(\sigma)I,$$
$$0 < \alpha_3(\sigma)I \leqq \tilde{\Phi}(t + \sigma, t)\tilde{W}(t, t + \sigma)\tilde{\Phi}'(t + \sigma, t) \leqq \alpha_4(\sigma) \exp [2m\sigma]I$$

holds for all t.

---

[3] The same result can be found in Kalman [2]; however, his proof contains errors.

[4] The same result is indicated by Anderson and Moore in [4] without proof.

5.4. Proof of Theorem 2. Assume that $R(A(t), B(t))$ is uniformly completely controllable. In a manner similar to the proof of Theorem 1, given any $m \geqq 0$, we transform $R(\hat{A}(t), 0)$ to

$$\dot{y}(t) = [A(t) + mI - B(t)K(t)]y(t), \tag{18}$$

where $y(t) = x(t) \exp [mt]$. Note that this is defined for all t (not necessarily for $t \geqq t_0$ as in the previous case).

Define $K(t)$ for any $m \geqq 0$ by

$$K(t) = \tfrac{1}{2}B'(t)Q_0(t), \tag{19}$$

where the $n \times n$ matrix $Q_0(t)$ is the solution of the matrix Riccati equation

$$\dot{Q}(t) + (A(t) + mI)'Q(t) + Q(t)(A(t) + mI) - Q(t)B(t)B'(t)Q(t) = -I.$$

The existence of $Q_0(t)$ is guaranteed by the assumption and Lemma 6 [2]. Since $Q_0(t)$ is continuous at every t, $K(t)$ defined by (19) is measurable and bounded on every finite subinterval of time.

Now, we shall show that $K(t)$ defined by (19) is a desired feedback gain. Consider a scalar function

$$V(y, t) = y'Q_0(t)y.$$

From the assumption, Lemma 6 and Lemma 5, $V(y, t)$ is bounded from above and below by nondecreasing functions of $\|y\|$ independent of t and $V(y, t) \to \infty$ as $\|y\| \to \infty$. The time derivative of $V(y, t)$ along any solution of (18) is nonpositive almost everywhere. Therefore, by means of Lyapunov's second method, there is a positive number a such that any solution of (18) satisfies

$$\|y(t_2)\| \leqq a\|y(t_1)\| \quad \text{for all } t_1 \text{ and } t_2 \geqq t_1,$$

or equivalently, any solution $x(t)$ of $R(\hat{A}(t), 0)$, where $K(t)$ is defined by (19), satisfies

$$\|x(t_2)\| \leqq a\|x(t_1)\| \exp [-m(t_2 - t_1)] \quad \text{for all } t_1 \text{ and } t_2 \geqq t_1.$$

5.5. Preliminary for the proof of Theorem 3. We make use of the following useful fact about a bounded system.

LEMMA 7 (Silverman and Anderson [12]). *A bounded representation $R(A(t), B(t))$ is uniformly completely controllable if and only if there exists a positive number σ such that*

$$0 < \alpha_1(\sigma)I \leqq W(t, t + \sigma) \quad \textit{for all } t.$$

5.6. Proof of Theorem 3.

Necessity. Let $\|A(t)\| \leqq K_1$. Then, by the Bellman–Gronwall inequality,

$$\|\Phi(t_2, t_1)\| \leqq \exp [K_1|t_2 - t_1|] \quad \text{for all } t_1 \text{ and } t_2.$$

Assume that bounded $R(A(t), B(t))$ is uniformly completely stabilizable by a bounded feedback gain. Let $\|K(t)\| \leqq K_2$. Then, since $\hat{\Phi}(\cdot, \cdot)$ satisfies

$$\hat{\Phi}(t_2, t_1) = \Phi(t_2, t_1) - \int_{t_1}^{t_2} \Phi(t_2, s)B(s)K(s)\hat{\Phi}(s, t_1)\, ds \quad \text{for all } t_1 \text{ and } t_2,$$

or equivalently,

$$\Phi(t_1, t_2)\hat{\Phi}(t_2, t_1) = I - \int_{t_1}^{t_2} \Phi(t_1, s)B(s)K(s)\hat{\Phi}(s, t_1)\, ds \quad \text{for all } t_1 \text{ and } t_2,$$

the following inequality holds for any n-vector μ:

$$a\|\mu\| \exp[(K_1 - m)(t_2 - t_1)] \geqq \|\mu\| - aK_2 \int_{t_1}^{t_2} \|\mu'\Phi(t_1, s)B(s)\|\, ds \tag{20}$$

for all t_1 and $t_2 \geqq t_1$.

If $R(A(t), B(t))$ is not uniformly completely controllable, then Lemma 7 implies that, for each $\sigma > 0$ and any $\rho > 0$, there is an n-vector λ ($\neq 0$) such that for some τ, $\lambda' W(\tau, \tau + \sigma)\lambda < \rho\lambda'\lambda$, or equivalently,

$$\int_{\tau}^{\tau+\sigma} \|\lambda'\Phi(\tau, s)B(s)\|^2\, ds < \rho\|\lambda\|^2. \tag{21}$$

Then, by Schwarz's inequality,

$$\int_{\tau}^{\tau+\sigma} \|\lambda'\Phi(\tau, s)B(s)\|\, ds < (\rho\sigma)^{1/2}\|\lambda\|.$$

Now we set $m = 2K_1$, $\sigma = (\log 3a)/K_1$, $\rho = (9a^2K_2^2\sigma)^{-1}$, and choose λ and τ which satisfy (21). Then, if we set $\mu = \lambda$, $t_1 = \tau$ and $t_2 = \tau + \sigma$ in (20),

$$\tfrac{1}{3}\|\lambda\| = a\|\lambda\| \exp[-K_1\sigma] \geqq \|\lambda\| - aK_2(\rho\sigma)^{1/2}\|\lambda\| = \tfrac{2}{3}\|\lambda\|.$$

This is a contradiction. Therefore, if bounded $R(A(t), B(t))$ is uniformly completely stabilizable by a bounded feedback gain, then it is uniformly completely controllable.

Sufficiency. The proof of Theorem 2 applies here. Moreover, since $B(t)$ is bounded in this case, $K(t)$ defined by (19) is bounded.

Remark 6. In the sufficiency part of the proofs of Theorems 1 and 2 different control laws and different Lyapunov functions are adopted for the following reason. In the proof of Theorem 1, the scalar function $y'P_0(t)y$ is not necessarily bounded from below by a nondecreasing function of $\|y\|$ which is independent of t (since $R(A(t), B(t))$ is only completely controllable). Thus the form $y'(P_0(t) + I)y$ is considered instead of $y'P_0(t)y$ as in Theorem 2.

REFERENCES

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

[2] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 102–119.

[3] G. W. JOHNSON, *A deterministic theory of estimation and control*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 380–384.

[4] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, N.J., 1971.
[5] W. A. WOLOVICH, *On the stabilization of controllable systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 569–572.
[6] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.
[7] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
[8] B. D. O. ANDERSON AND J. B. MOORE, *Linear system optimisation with prescribed degree of stability*, Proc. IEE, 116 (1969), pp. 2083–2087.
[9] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1962), pp. 189–213.
[10] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.
[11] L. A. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.
[12] L. M. SILVERMAN AND B. D. O. ANDERSON, *Controllability, observability and stability of linear systems*, this Journal, 6 (1968), pp. 121–130.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

A NOTE ON THE PENALTY METHOD FOR DISTRIBUTED PARAMETER OPTIMAL CONTROL PROBLEMS*

HITOSHI SASAI†

Abstract. In this paper we extend A. V. Balakrishnan, A. P. Jones and G. P. McCormick's penalty formulation to optimal control problems described by partial differential equations, e.g., linear distributed parameter systems with inequality constraints.

1. Introduction. The application of the penalty method for optimal control problems subject to side constraints is currently of much interest.

The original formulation of the penalty method is due to R. Courant [1]. Further development of this study was made by A. V. Fiacco and G. P. McCormick [2], and A. V. Fiacco and A. P. Jones [3].

The extensions of the penalty method to the field of optimal control problems described by ordinary differential equations can be found in [4] and [5].

Recently, A. V. Balakrishnan [6] has developed a new idea of viewing the system equations as the equality constraints, thereby avoiding solving the system equations.

Another penalty formulation, which is a generalization of Balakrishnan's and uses a mixed interior penalty function, is made by A. P. Jones and G. P. McCormick [7]. They approach the optimum of constrained problems by generating a sequence of minimizing points of penalty functions. The application of this method is limited to the problems of ordinary differential equations.

In this paper we extend the work of Balakrishnan [6] and Jones and McCormick [7] by allowing more general equations relating the state and control unknowns. We extend their penalty methods to problems described by partial differential equations, e.g., inequality constrained linear distributed parameter optimal control problems.

2. Statement of problem. Let H_1 and H_2 be Hilbert spaces. We consider the following distributed parameter system:

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = 0, \tag{1}$$

where $x(t)$ is an element of H_1 for each t, $0 \leqq t < \infty$, $u(t)$ is an element of H_2 for each t, $0 \leqq t < \infty$, and is square integrable in t, A is an unbounded linear operator, mapping a domain $D(A)$ dense in H_1 into H_1, and B is a linear bounded operator mapping H_2 into H_1. We use the same symbol for both the norms in H_1 and H_2.

Let $B_2(H_1, T)$ and $B_2(H_2, T)$ be the spaces of strongly measurable functions $x(t)$ and $u(t)$ with range in H_1 and H_2 such that

$$\int_0^T |x(t)|^2 \, dt < \infty \quad \text{and} \quad \int_0^T |u(t)|^2 \, dt < \infty,$$

* Received by the editors May 14, 1971, and in final revised form December 6, 1971.

† Department of Aeronautical Engineering, Nagoya University, Chikusa-ku Nagoya 464, Japan.

respectively. Moreover, we define the norms $\|x(t)\|$ and $\|u(t)\|$ of each space by

$$\|x(t)\|^2 = \int_0^T |x(t)|^2 \, dt \quad \text{and} \quad \|u(t)\|^2 = \int_0^T |u(t)|^2 \, dt,$$

respectively. We take the solution of (1) in the space $B_2(H_1, T)$.

Now we introduce the operator S defined over a suitable dense subset $D(S)$ of $B_2(H_1, T)$ by

$$S = \partial/\partial t - A,$$

where we assume that A is such that S is a closable one and let $\bar{S}$ be the minimal closed extension of S. We say that $x(t) \in D(S)$ is a solution of (1) for $u(t)$ if the following is satisfied:

$$(x(t), (S^*\varphi)(t)) = (Bu(t), \varphi(t)) \quad \text{for all } \varphi(t) \in D(S^*), \tag{2}$$

where $(\cdot, \cdot)$ denotes the inner product in $B_2(H_1, T)$ and S^* is the adjoint of S.

Remark. Since $D(S)$ is dense in $B_2(H_1, T)$, S^* always exists. Moreover, $D(S^*)$ is dense in $B_2(H_1, T)$ and $S^{**} = \bar{S}$, because S is closable and we are dealing with Hilbert spaces. Therefore (2) is equivalent to

$$\bar{S}x(t) = Bu(t),$$

according to [8].

Now we consider the following problems.

Problem (A). Find $u(t) \in B_2(H_2, T)$ minimizing

$$\int_0^T f(x(t), u(t)) \, dt$$

subject to the equation (2) and the inequalities

$$g_i(x(t), u(t)) \geqq 0, \qquad i = 1, \cdots, m, \quad \text{for a.a.t,}$$

where f and g_i satisfy the following condition:

(C_1): $f(v, w): H_1 \times H_2 \to R_1$ is continuous in each variable v and w, convex and bounded on bounded sets. The $g_i(v, w): H_1 \times H_2 \to R_1$ are continuous in each variable and concave.

For the next problem, we define

$$Q \equiv \{x(t) \in D(\bar{S}), u(t) \in B_2(H_2, T) | \|(\bar{S}x)(t) - Bu(t)\| = 0\},$$

$$R \equiv \{x(t) \in B_2(H_1, T), u(t) \in B_2(H_2, T) | g_i(x(t), u(t)) \geqq 0,$$
$$i = 1, \cdots, m, \text{ for a.a.t}\},$$

$$R^0 \equiv \{x(t) \in B_2(H_1, T), u(t) \in B_2(H_2, T) | g_i(x(t), u(t)) > 0,$$
$$i = 1, \cdots, m, \text{ for a.a.t, and } \int_0^T \frac{1}{g_i(x(t), u(t))} \, dt < \infty\},$$

$$\Omega \equiv \{v \in H_1, w \in H_2 | g_i(v, w) \geqq 0, i = 1, \cdots, m\};$$

(C_2): there exists $(x(t), u(t)) \in R^0$, where $x(t) \in D(\bar{S})$;

(C_3): Ω is bounded;

(C_4): $f(v, w)$ is bounded from below.

Problem (B). Find $x(t)$ and $u(t)$ minimizing

$$P(x(t), u(t), r_k) \equiv \int_0^T f(x(t), u(t))\, dt + r_k \sum \int_0^T \frac{1}{g_i(x(t), u(t))}\, dt + \int_0^T \frac{|(\bar{S}x)(t) - Bu(t)|^2}{r_k}\, dt, \tag{3}$$

over $(x(t), u(t)) \in R$ and $x(t) \in D(\bar{S})$, for each $r_k > 0$, where $r_k \to 0$ as $k \to \infty$ and $r_k > r_{k+1}$.

In the sequel we get the computational solution of Problem (A) by showing that the solution $(x_{r_k}(t), u_{r_k}(t))$ of Problem (B) exists for every $r_k > 0$, and it converges to the solution of Problem (A) as $r_k \to 0$.

LEMMA 1. *Under* (C_1) *and* (C_3) R *is closed, convex and bounded and* Q *is closed and convex in* $B_2(H_1, T) \times B_2(H_2, T)$. *Therefore* $R \cap Q$ *is closed, convex and bounded.*

Proof. It is clear that by (C_1), R and Q are convex, and that by (C_3) and (C_1), R is bounded and closed.

Next we shall show that Q is closed. Since $\bar{S}$ is the closed operator,

$$\{(x(t), Bu(t)) \| \|(\bar{S}x)(t) - Bu(t)\| = 0\} = (x(t), (\bar{S}x)(t))$$

is closed in $B_2(H_1, T) \times B_2(H_1, T)$. Therefore Q is closed because B is a bounded operator.

Remark. The convexity of R and Q implies their weak closedness and hence also for $R \cap Q$.

THEOREM 1. *Suppose that* (C_1), (C_2), (C_3) *and* (C_4) *are satisfied. Then the function* $\int_0^T f(x(t), u(t))\, dt$ *takes a minimum over* $R \cap Q$.

Proof. Since $R \cap Q$ is closed, convex and bounded by Lemma 1 (i.e., weakly closed), and $\int_0^T f(x(t), u(t))\, dt$ is weakly lower semicontinuous [9], there exists finite v_0 such that

$$v_0 = \inf_{R \cap Q} \int_0^T f(x(t), u(t))\, dt = \min_{R \cap Q} \int_0^T f(x(t), u(t))\, dt = \int_0^T f(x^*(t), u^*(t))\, dt.$$

Remark. This theorem means that a solution of Problem (A) exists, i.e., an optimal control. $u^*(t)$ is an optimal control and $x^*(t)$ is the optimal trajectory corresponding to $u^*(t)$.

Now let $x_0(t) \in D(\bar{S})$ and $(x_0(t), u_0(t)) \in R^0$ and define $M_0 \equiv P(x_0(t), u_0(t), r_k)$. Put

$$V_0 \equiv \left\{ x(t), u(t) \middle| \int_0^T f(x(t), u(t))\, dt + \int_0^T \frac{|(\bar{S}x)(t) - Bu(t)|^2}{r_k}\, dt \leqq M_0, \ (x(t), u(t)) \in R \right\}.$$

It can be shown that there exists a finite δ_k (see the Appendix) such that

$$\delta_k = \min_{V_0} \left\{ \int_0^T f(x(t), u(t))\, dt + \int_0^T \frac{|(\bar{S}x)(t) - Bu(t)|^2}{r_k}\, dt \right\}.$$

We define V_1, V as follows:

$$V_1 \equiv \left\{ x(t), u(t) \middle| \int_0^T \frac{r_k}{g_i(x(t), u(t))}\, dt \leqq M_0 - \delta_k, i = 1, \cdots, m \right\},$$

$$V = V_0 \cap V_1$$

(V is not empty, for $(x_0(t), u_0(t)) \in V$).

By a technique similar to that in the Appendix, the following lemma can be easily proved.

LEMMA 2. *$P(x(t), u(t), r_k)$ attains a minimum over V for each r_k.*

LEMMA 3. *If $(x(t), u(t)) \in R$ and $P(x(t), u(t), r_k) \leqq P(x_0(t), u_0(t), r_k) = M_0$, then $(x(t), u(t)) \in V$.*

Proof. If $(x(t), u(t)) \in R$ and $P(x(t), u(t), r_k) \leqq M_0$, it follows that $(x(t), u(t)) \in V_0$ by the following inequality:

$$M_0 \geqq P(x(t), u(t), r_k) \geqq \int_0^T f(x(t), u(t))\, dt + \int_0^T \frac{|(\bar{S}x)(t) - Bu(t)|^2}{r_k}\, dt.$$

Next we shall show that $(x(t), u(t)) \in V_1$. Let us assume $(x(t), u(t)) \notin V_1$, i.e.,

$$\int_0^T \frac{r_k}{g_j(x(t), u(t))}\, dt > M_0 - \delta_k \quad \text{for some } i = j.$$

Since $(x(t), u(t)) \in V_0$,

$$\begin{aligned} P(x(t), u(t), r_k) &= \int_0^T f(x(t), u(t))\, dt + \int_0^T \frac{r_k}{g_j(x(t), u(t))}\, dt \\ &\quad + \sum_{i \neq j} \int_0^T \frac{r_k}{g_i(x(t), u(t))}\, dt + \int_0^T \frac{|(\bar{S}x)(t) - Bu(t)|^2}{r_k}\, dt \\ &> \int_0^T f(x(t), u(t))\, dt + \int_0^T \frac{|(\bar{S}x)(t) - Bu(t)|^2}{r_k}\, dt \\ &\quad + M_0 - \delta_k > M_0, \end{aligned}$$

which contradicts the assumptions of Lemma 3.

Thus, we have $(x(t), u(t)) \in V$.

THEOREM 2. *Suppose that (C_1), (C_2), (C_3) and (C_4) are satisfied. Then, for each r_k, Problem (B) has a solution $(x_{r_k}(t), u_{r_k}(t))$ which is in R^0 (i.e., $P(x(t), u(t), r_k)$ takes a minimum at $(x_{r_k}(t), u_{r_k}(t)) \in R^0$).*

Proof. This theorem is clear by Lemma 2 and Lemma 3, because $V \subseteq R^0$.

We now assume (see [7]):

(C_5): for any $\varepsilon > 0$ there exists $(x(t), u(t)) \in R^0 \cap Q$ such that

$$\int_0^T f(x(t), u(t))\, dt < \inf_{R \cap Q} \int_0^T f(x(t), u(t))\, dt + \varepsilon = v_0 + \varepsilon.$$

THEOREM 3. *Suppose that* (C_1), (C_2), (C_3), (C_4) *and* (C_5) *are satisfied. Then we have*:

$$\lim_{k\to\infty} P(x_{r_k}(t), u_{r_k}(t), r_k) = \int_0^T f(x^*(t), u^*(t))\,dt = v_0.$$

Proof. By (C_5) there exists $(x(t), u(t)) \in R^0 \cap Q$ such that

$$\int_0^T f(x(t), u(t))\,dt < v_0 + \frac{\varepsilon}{2}.$$

Following [7] we select k' such that

$$\max_i \int_0^T \frac{r_{k'}}{g_i(x(t), u(t))}\,dt < \frac{\varepsilon}{2m}.$$

Then for $k > k'$ we have

$$P(x_{r_k}(t), u_{r_k}(t), r_k) \leqq P(x(t), u(t), r_{k'}) < v_0 + \varepsilon. \tag{4}$$

By (4), each term of $P(x_{r_k}(t), u_{r_k}(t), r_k)$ is bounded, for each term of $P(x_{r_k}(t), u_{r_k}(t), r_k)$ is bounded below.

Thus by the technique similar to that in the Appendix, there exist subsequences $x_{r_{kj}}(t), u_{r_{kj}}(t)$ and $(\bar{S}x_{r_{kj}})(t)$ of $x_{r_k}(t), u_{r_k}(t)$ and $(\bar{S}x_{r_k})(t)$ which converge weakly to $\hat{x}(t)$, $\hat{u}(t)$ and $(\bar{S}\hat{x})(t)$, respectively.

Since $\int_0^T |(\bar{S}x_{r_{kj}})(t) - Bu_{r_{kj}}(t)|^2\,dt/r_{kj}$ is bounded, we have

$$\|(\bar{S}\hat{x})(t) - B\hat{u}(t)\| \leqq \liminf \|(\bar{S}x_{r_{kj}})(t) - Bu_{r_{kj}}(t)\| = 0,$$

i.e., $(\hat{x}(t), \hat{u}(t)) \in Q$ (see [7]).

Since R is weakly closed, $(\hat{x}(t), \hat{x}(t)) \in R \cap Q$ and hence we have:

$$v_0 \leqq \int_0^T f(\hat{x}(t), \hat{u}(t))\,dt \leqq \liminf \int_0^T f(x_{r_{kj}}(t), u_{r_{kj}}(t))\,dt$$

$$\leqq \lim P(x_{r_{kj}}(t), u_{r_{kj}}(t), r_{kj}) \leqq v_0 + \varepsilon,$$

because $\int_0^T f(x(t), u(t))\,dt$ is weakly lower semicontinuous.

Consequently,

$$\lim_{j\to\infty} P(x_{r_{kj}}(t), u_{r_{kj}}(t), r_{kj}) = v_0.$$

Any subsequence of $P(x_{r_k}(t), u_{r_k}(t), r_k)$ has its subsequence converging to v_0, and therefore $P(x_{r_k}(t), u_{r_k}(t), r_k)$ converges to v_0.

Remark. By the results in [2, p. 819] and in [6, p. 162], we see the following:

$$\lim \int_0^T f(x_{r_k}(t), u_{r_k}(t))\,dt = v_0,$$

$$\lim \int_0^T \frac{r_k}{g_i(x_{r_k}(t), u_{r_k}(t))}\,dt = 0,$$

$$\lim \int_0^T \frac{|(\bar{S}x_{r_k})(t) - Bu(t)|^2}{r_k}\,dt = 0.$$

3. Conclusion. In this paper we have extended the penalty method which was developed in [7] to the problems described by partial differential equations, e.g., inequality constrained distributed parameter optimal control problems.

We have shown under stated conditions that the unconstrained problems (Problem (B)) have solutions interior to the constraint sets (Theorem 2) and in the limit these solutions converge to the constrained optimum, i.e., solve Problem (A) (Theorem 3).

Appendix. *The function*

$$h(x(t), u(t)) \equiv \int_0^T f(x(t), u(t))\, dt + \int_0^T |(\bar{S}x)(t) - Bu(t)|^2 / r_k\, dt$$

attains a finite minimum over V_0.

Proof. We put $d \equiv \inf_{V_0} h(x(t), u(t))$ and let $(x_n(t), u_n(t))$ be a minimizing sequence tending to d.

$R \supseteq V_0$ being bounded, $x_n(t)$ and $u_n(t)$ are bounded in $B_2(H_1, T)$ and $B_2(H_2, T)$ respectively. Hence $(\bar{S}x_n)(t)$ is bounded, because f is bounded below and B is the bounded operator. Therefore there exist subsequences of $x_n(t), u_n(t)$ and $(\bar{S}x_n)(t)$ which converge weakly to $\hat{x}(t), \hat{u}(t)$ and $\hat{y}(t)$, respectively. We write n for the subsequence n_j.

It can be shown that $\hat{x}(t) \in D(\bar{S})$ and $(\bar{S}\hat{x})(t) = \hat{y}(t)$, because the following equalities are satisfied:

$$[\hat{y}(t), \varphi(t)] = \lim [(\bar{S}x_n)(t), \varphi(t)] = \lim [x_n(t), (S^*\varphi)(t)]$$
$$= [\hat{x}(t), (S^*\varphi)(t)] \quad \text{for } \varphi(t) \in D(S^*).$$

Now we can conclude that $(\hat{x}(t), \hat{u}(t)) \in V_0$ and $d = h(\hat{x}(t), \hat{u}(t))$ because $h(x(t), u(t))$ is weakly lower semicontinuous and R is weakly closed.

Acknowledgment. The author wishes to thank Professor H. Ishigaki, the School of Education, Waseda University, for helpful suggestions, and Professor I. Sugiura, Nagoya University, and Professor E. Shimemura, Waseda University, for their encouragement during the preparation of this paper.

REFERENCES

[1] R. COURANT, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.

[2] A. V. FIACCO AND G. P. MCCORMICK, *Extensions of SUMT for nonlinear programming*, Management Sci., 12 (1966), pp. 816–828.

[3] A. V. FIACCO AND A. P. JONES, *Generalized penalty methods in topological spaces*, SIAM J. Appl. Math., 5 (1969), pp. 317–331.

[4] L. S. LASDON, A. D. WAREN AND R. K. RICE, *An interior penalty method for inequality constrained optimal control problems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 388–395.

[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. BALAKRISHNAN, *On a new computing method in optimal control*, this Journal, 6 (1968), pp. 149–173.

[7] A. P. JONES AND G. P. MCCORMICK, *A generalization of the method of Balakrishnan: Inequality constraints and initial conditions*, this Journal, 8 (1970), pp. 218–225.

[8] S. DE JULIO, *On the optimization of infinite dimensional linear systems*, Proc. 2nd International Conference on Computing Methods in Optimization Problems, San Remo, Italy, 1968.

[9] P. VARAIYA, *N-person nonzero sum differential games with linear dynamics*, this Journal, 8 (1970), pp. 441–449.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

CONVOLUTION FEEDBACK SYSTEMS*

C. A. DESOER AND F. M. CALLIER†

Abstract. This paper considers multi-input multi-output feedback systems characterized by $y = G * e$ and $e = u - y$. Theorem 1 shows that if the closed loop impulse response H is stable in the sense that $H \in \mathscr{A}^{n\times n}(\sigma)$, then $\hat{G}(s) = \hat{P}(s)[\hat{Q}(s)]^{-1}$, where $\hat{P}(s)$, $\hat{Q}(s)$ are also in $\mathscr{A}^{n\times n}(\sigma)$. Theorem 2 gives necessary and sufficient conditions for $H \in \mathscr{A}^{n\times n}(\sigma)$. Finally Theorem 3 gives necessary and sufficient conditions for stability when $\hat{G}(s)$ has a multiple pole in $\operatorname{Re} s > \sigma$; the case where the leading term of the Laurent expansion at this pole is singular is treated in detail. The case of a finite number of multiple poles follows easily.

1. Introduction. This paper considers linear time-invariant feedback systems with n inputs and n outputs. As will become apparent, there is no loss of generality in taking the feedback to be unity. The input u, output y and error e are functions from $\mathbb{R}_+$ (defined as $[0, \infty)$) to $\mathbb{R}^n$ or corresponding distributions on $\mathbb{R}_+$. The open loop system is of the convolution type so that we have

$$y = G * e, \tag{1}$$

$$e = u - y. \tag{2}$$

G is an $n \times n$ matrix whose elements are distributions on $\mathbb{R}_+$. We use $\underline{G}$ to denote the map $\underline{G}: e \mapsto G * e$.

We shall repeatedly use the convolution algebra $\mathscr{A}(\sigma)$ (see [1], [2]): f is said to be in $\mathscr{A}(\sigma)$ if and only if $f(t) = 0$ for $t < 0$ and

$$f(t) = f_a(t) + \sum_0^\infty f_i \delta(t - t_i) \quad \text{for } t \geqq 0, \tag{3}$$

where $f_a(t)\, e^{-\sigma t} \in L^1(0, \infty)$, $f_i \in \mathbb{R}$ for all i, $\sum_0^\infty |f_i|\, e^{-\sigma t_i} < \infty$ and $0 = t_0 < t_1 < t_2 < \cdots$. Thus f is a distribution of order 0 with support on $\mathbb{R}_+$. An n-vector v ($n \times n$ matrix A) is said to be in $\mathscr{A}^n(\sigma)$ ($\mathscr{A}^{n\times n}(\sigma)$) if and only if all its elements are in $\mathscr{A}(\sigma)$. Let $\hat{f}$ denote the Laplace transform of f: f belongs to the convolution algebra $\mathscr{A}(\sigma)$ if and only if $\hat{f}$ belongs to the algebra $\hat{\mathscr{A}}(\sigma)$ (with pointwise product). Similarly, $\hat{v} \in \hat{\mathscr{A}}^n(\sigma)$, $\hat{A} \in \hat{\mathscr{A}}^{n\times n}(\sigma)$.

Recently M. Vidyasagar [5] has shown that the class of systems (1), (2) where

$$\hat{G}(s) = \hat{P}(s)[\hat{Q}(s)]^{-1} \tag{4}$$

with $\hat{P}, \hat{Q} \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ is very useful for distributed networks, for example, and he extended some stability results of Desoer, Wu, Baker, Vakharia and Lam [1]–[3], [10]. In Theorem 1 below we prove that, under very mild assumptions on G and on the closed loop system, if the closed loop impulse response $H \in \mathscr{A}^{n\times n}(\sigma)$, then $\hat{G}$ is of the form (4). Theorem 1 is also an extension of a result of Nasburg and Baker [4]: the extension is in two directions, first, the n-input n-output case

* Received by the editors August 5, 1971, and in revised form December 6, 1971.

† College of Engineering, Electronics Research Laboratory, University of California, Berkeley, California 94720. This research was supported by the National Aeronautics and Space Administration under Grant NGL-05-003-016.

is considered and, second, the requirements on G are greatly relaxed. Theorem 2 is a straightforward extension of a result of [4]: it shows the importance of the systems considered by Vidyasagar in the sense that $H \in \mathscr{A}^{n\times n}(\sigma)$ if and only if $\hat{G}$ is of the form (4). Finally, Theorem 3 gives the *necessary* and *sufficient* conditions for stability of the closed loop system when $\hat{G}$ is of the form (4) with a finite number of poles of finite order in $\operatorname{Re} s > \sigma$. This theorem culminates a series of investigations starting with [1]–[3], [7]. Note that except for [7], all previous work could only prove sufficiency.

2. The relation between G and H.

THEOREM 1. *Let G be an $n \times n$ matrix whose elements are distributions with support on $\mathbb{R}_+$. Suppose that in a neighborhood of the origin, say $V \subset \mathbb{R}$, G includes at most δ-functions (i.e., on V, it is a distribution of at most order 0). For the system defined by (1) and (2), assume that the closed loop response H exists and is uniquely defined by*

$$H + G * H = G. \tag{5}$$

Under these conditions, if $H \in \mathscr{A}^{n\times n}(\sigma)$, then

(i) *G is Laplace transformable and for some $\bar{\sigma} \geqq \sigma$, $G \in \mathscr{A}^{n\times n}(\bar{\sigma})$.*

(ii) *$\hat{G}$ is of the form*

$$\hat{G}(s) = \hat{P}(s)[\hat{Q}(s)]^{-1} \quad \text{for} \quad \operatorname{Re} s > \sigma, \tag{6}$$

where $\hat{P}(\cdot)$ and $\hat{Q}(\cdot) \in \mathscr{A}^{n\times n}(\sigma)$.

(iii) *$\hat{G}$ can at most have a countable number of poles in $\operatorname{Re} s > \sigma$.*

Comment. This theorem shows that under mild conditions on G regarding its behavior near $t = 0$, once the closed loop system is well-defined and "stable", then $\hat{G}$ is *necessarily* of the form (6), can at most have poles in the strip $\sigma < \operatorname{Re} s \leqq \bar{\sigma}$ and is analytic for $\operatorname{Re} s > \bar{\sigma}$.

Proof. (i) By assumption, H is of the form

$$H(t) = H_a(t) + \sum_{i=0}^{\infty} H_i \delta(t - t_i),$$

where $0 = t_0 < t_1 < t_2 < \cdots$. By assumption G can at most have an impulse at the origin. By the Abelian theorem of the Laplace transform [11] and the properties of distributions, if G has an impulse G_0 at $t = 0$, $\hat{G}(s) \to G_0$ as $s \to \infty$ with $\operatorname{Re} s \to \infty$. Clearly from (5), if G_0 is the zero matrix, then $H_0 = 0$. If $G_0 \neq 0$, then by balancing impulses at the origin in (5) we have $(I + G_0)H_0 = G_0$. By assumption H, hence H_0, is uniquely defined by (5), hence $\det(I + G_0) \neq 0$. Furthermore, by direct calculation, $(I + G_0)(I - H_0) = I$, so that $\det[I - H_0] \neq 0$.

The function $I - \hat{H}(s)$ is analytic and bounded for $\operatorname{Re} s > \sigma$, and tends to $I - H_0$ as $s \to \infty$ with $\operatorname{Re} s \to \infty$. Consequently, there exists a $\bar{\sigma} \geqq \sigma$ such that

$$\inf_{\operatorname{Re} s \geqq \bar{\sigma}} |\det[I - \hat{H}(s)]| > 0. \tag{7}$$

From (5), if G had a Laplace transform, we would have $\hat{H} + \hat{G}\hat{H} = \hat{G}$. Now by

(7), $\hat{H}(s)[I - \hat{H}(s)]^{-1} \in \hat{\mathscr{A}}^{n\times n}(\bar{\sigma})$, so $\hat{G}(s)$ is equal to that function, by the uniqueness of the convolution algebra of distributions on $\mathbb{R}_+$.

(ii) Since $\hat{H}(s)$ is analytic for $\operatorname{Re} s > \sigma$, $[I - \hat{H}(s)]^{-1}$ has at most a countable number of poles in $\operatorname{Re} s > \sigma$ and by analytic continuation,

$$\hat{G}(s) = \hat{H}(s)[I - \hat{H}(s)]^{-1} \quad \text{for} \quad \operatorname{Re} s > \sigma. \tag{8}$$

Choose $\hat{P}(s) = \hat{H}(s)$, $\hat{Q}(s) = I - \hat{H}(s)$. Thus (ii) and (iii) have been established.

Remark. It is important to reflect on the fact that under the conditions of Theorem 1, we have

$$[I + \hat{G}(s)][I - \hat{H}(s)] = I \quad \text{for} \quad \operatorname{Re} s > \sigma.$$

This expression emphasizes the symmetrical role played by H and G: H is obtained from G by a negative feedback of I; G is obtained from H by a negative feedback of $-I$ (to cancel the preceding one!).

THEOREM 2. *Let G be an $n \times n$ matrix whose elements are Laplace transformable distributions with support in $\mathbb{R}_+$. For the system defined by* (1) *and* (2), *assume that the closed loop transfer function $\hat{H}$ is well-defined for almost all s in the half-plane of convergence of $\hat{G}$; i.e.,*

$$\hat{H}(s) = \hat{G}(s)[I + \hat{G}(s)]^{-1}; \tag{9}$$

thus $\hat{H}(\cdot)$ is meromorphic in the (open) half-plane of convergence of $\hat{G}(\cdot)$. Under these conditions,

$$H \in \mathscr{A}^{n\times n}(\sigma) \quad \textit{for some } \sigma \tag{10}$$

if and only if there exist $\hat{P}, \hat{Q} \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ such that

$$\hat{G}(s) = \hat{P}(s)[\hat{Q}(s)]^{-1} \quad \textit{for} \quad \operatorname{Re} s > \sigma \tag{11}$$

and

$$\inf_{\operatorname{Re} s \geq \sigma} |\det [\hat{P}(s) + \hat{Q}(s)]| > 0. \tag{12}$$

Proof. Necessity. From (9) by algebra,

$$\hat{G}(s) = \hat{H}(s)[I - \hat{H}(s)]^{-1} \quad \text{for} \quad \operatorname{Re} s > \sigma.$$

Choose $\hat{P}(s) = \hat{H}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ and $\hat{Q}(s) = I - \hat{H}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$, by (10). Since $\hat{P} + \hat{Q} = I$, (12) holds.

Sufficiency. From (9) and (11),

$$\hat{H}(s) = \hat{P}(s)[\hat{P}(s) + \hat{Q}(s)]^{-1}.$$

In view of (12), $\hat{H} \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ as the product of two elements of $\hat{\mathscr{A}}^{n\times n}(\sigma)$.

Remark. It is clear from (11) that a given $\hat{G}$ does not define the ordered pair $(\hat{P}, \hat{Q})$ uniquely; for example, they might have a matrix as right common factor. In order to be able to express the condition (12) in a form which depends on $\hat{G}$ only, we impose the Vidyasagar no-cancellation condition (N) [5]: the ordered pair (a, b), where $a, b: \mathbb{C} \to \mathbb{C}$ is said to satisfy the no-cancellation condition on a set $A \subset \mathbb{C}$ if and only if, for all sequences $\{s_k\}$ in A, $a(s_k) \to 0$ implies that $\liminf |b(s_k)| > 0$.

It is then easy to show [5] that if $(\det \hat{Q}(s), \det[\hat{P}(s) + \hat{Q}(s)])$ satisfies (N) on $\operatorname{Re} s \geqq \sigma$, then (12) is equivalent to $\inf_{\operatorname{Re} s \geqq \sigma} |\det[I + \hat{G}(s)]| > 0$.

3. Necessary and sufficient conditions for stability. We consider first and in detail the case where $\hat{G}$ has a single pole p of order m in $\operatorname{Re} s > \sigma$. The extension to the case of a finite number of poles is straightforward.

We consider the open loop transfer function

$$\hat{G}(s) = \sum_{i=0}^{m-1} R_i(s-p)^{-m+i} + \hat{G}_0(s), \tag{13}$$

where $\operatorname{Re} p > \sigma$, $\hat{G}_0 \in \hat{\mathscr{A}}^{n\times n}(\sigma)$, $r_0 \triangleq \operatorname{rank} R_0 \leqq n$ and R_i, $i = 0, 1, \cdots, m-1$, are $n \times n$ matrices with complex coefficients. We start by pointing out some facts which will streamline the proof.

Fact 1. Let

$$\hat{\tilde{R}}\left(\frac{1}{s+a}\right) \triangleq \left(\sum_{i=0}^{m-1} R_i(s-p)^{-m+i}\right)\left(\frac{s-p}{s+a}\right)^m, \quad a \triangleq 1-\sigma; \tag{14}$$

then $\hat{\tilde{R}}(1/s+a))$ is an $n \times n$ complex *polynomial* matrix in $(1/(s+a))$ of degree m. This is obvious by considering the Laurent expansion of $\hat{\tilde{R}}(1/(s+a))$ about $s = -a$.

Fact 2 (Smith canonical form [12]). For the $n \times n$ polynomial matrix $\hat{\tilde{R}}(1/(s+a))$ there exist *unimodular* (i.e., with nonzero *constant* determinant) *polynomial* matrices in $(1/(s+a))$, viz., $\hat{P}(1/(s+a))$ and $\hat{Q}(1/(s+a))$, such that:

$$\hat{Q}\left(\frac{1}{s+a}\right)\hat{\tilde{R}}\left(\frac{1}{s+a}\right)\hat{P}\left(\frac{1}{s+a}\right) = \operatorname{diag}\Bigg\{\underbrace{\hat{a}_1\left(\frac{1}{s+a}\right), \cdots, \hat{a}_j\left(\frac{1}{s+a}\right), \cdots, \hat{a}_r\left(\frac{1}{s+a}\right)}_{r}, \underbrace{0, 0, \cdots, 0}_{n-r}\Bigg\}, \tag{15}$$

where

(i) r = rank of $\hat{\tilde{R}}(1/(s+a))$ = order of the largest minor of $\hat{\tilde{R}}(1/(s+a))$ which is not equal to the zero polynomial;

(ii) $\hat{a}_j(1/(s+a)), j = 1, 2, \cdots, r$, are the invariant polynomials of $\hat{\tilde{R}}(1/(s+a))$ and each polynomial $\hat{a}_j(\cdot)$ divides $\hat{a}_{j+1}(\cdot), j = 1, 2, \cdots, r-1$;

(iii) the diagonal matrix in the right-hand side of (15) can be obtained by elementary operations.

Fact 3. The polynomial matrices $\hat{P}(1/(s+a))$ and $\hat{Q}(1/(s+a)) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ and their inverses are polynomial matrices in $(1/(s+a))$, also in $\hat{\mathscr{A}}^{n\times n}(\sigma)$.

Fact 4. Let $\hat{a}_j(\cdot), j = 1, 2, \cdots, r$, be as in (15) and let r_0 be the rank of R_0. Then

(a)

$$\begin{aligned} &\hat{a}_j(1/(p+a)) = 0 \quad \text{for} \quad r_0 + 1 \leqq j \leqq r \text{ by definition of } r_0, \\ &\hat{a}_j(1/(p+a)) \neq 0 \quad \text{for} \quad 1 \leqq j \leqq r_0; \end{aligned} \tag{16}$$

(b)

$$\hat{a}_j\left(\frac{1}{s+a}\right) = \hat{b}_j\left(\frac{1}{s+a}\right)\left(\frac{s-p}{s+a}\right)^{c_j} \quad \text{for} \quad r_0 + 1 \leqq j \leqq r, \tag{17}$$

where c_j is the order of the zero of $\hat{a}_j(\cdot)$ at $s = p$;
$\hat{b}_j(\cdot)$ is a polynomial with

$$\hat{b}_j(1/(p+a)) \neq 0 \qquad \text{(see [13])} \tag{18}$$

and $1 \leqq c_{r_0+1} \leqq c_{r_0+2} \leqq \cdots \leqq c_r$.

Proof. Set $s = p$ in (15) and note that the left-hand side becomes $\hat{Q}(1/(p+a))$ $R_0(p+a)^{-m}\hat{P}(1/(p+a))$. Since $\hat{P}(\cdot)$ and $\hat{Q}(\cdot)$ are unimodular, exactly $r - r_0$ polynomials $\hat{a}_j(\cdot)$ are zero at $s = p$. By (ii) of (15), $\hat{a}_j(1/(p+a)) = 0$ for $r_0 + 1 \leqq j \leqq r$. Hence (16) and (17) follow with the properties of the latter as a consequence of (ii) of (15).

Remark. Note that the exponents c_j in (17) may, for some j, be larger than m (in fact $c_r \leqq rm$).

Therefore, since the c_j are monotonically increasing and since $c_j - m$ may be of any sign, partition the index set $K = \{r_0 + 1, r_0 + 2, \cdots, r\}$ into

$$K_- = \{r_0 + 1, r_0 + 2, \cdots, \alpha\} = \{j | 1 \leqq c_j < m\}, \tag{19}$$

$$K_0 = \{\alpha + 1, \alpha + 2, \cdots, \beta\} = \{j | c_j = m\}, \tag{20}$$

$$K_+ = \{\beta + 1, \beta + 2, \cdots, r\} = \{j | c_j > m\}. \tag{21}$$

We are now ready for Theorem 3.

THEOREM 3. *Let $\hat{G}(s)$ be given by* (13) *and let $\hat{P}(1/(s+a))$ and $\hat{Q}(1/(s+a))$ be the polynomial matrices defined in* (15). *Suppose that the index sets K_-, K_0, K_+, as defined in* (19)–(21), *are not empty.*

Consider the partitioning

$$\hat{Q}\left(\frac{1}{s+a}\right)[I + \hat{G}_0(s)]\hat{P}\left(\frac{1}{s+a}\right) = \begin{array}{c} \\ \alpha \\ n-\alpha \end{array}\!\!\begin{array}{c} \overbrace{\phantom{\hat{L}_{11}(s)}}^{\alpha} \quad \overbrace{\phantom{\hat{L}_{12}(s)}}^{n-\alpha} \\ \left[\begin{array}{c|c} \hat{L}_{11}(s) & \hat{L}_{12}(s) \\ \hline \hat{L}_{21}(s) & \hat{L}_{22}(s) \end{array}\right] \end{array} \tag{22}$$

and let $\hat{b}_j(\cdot)$ be the polynomials defined in (17). *Under these conditions,*

$$H \in \mathscr{A}^{n\times n}(\sigma)$$

if and only if

$$\inf_{\operatorname{Re} s \geqq \sigma} |\det [I + \hat{G}(s)]| > 0 \tag{23}$$

and

(C) $\det\{\hat{L}_{22}(p) + \operatorname{diag}[\hat{b}_{\alpha+1}(1/(p+a)), \cdots, \hat{b}_\beta(1/(p+a)), 0, 0, \cdots, 0]\} \neq 0.$

Proof. Sufficiency. Since $I - \hat{H}(s) = [I + \hat{G}(s)]^{-1}$, we need only to show that

$$[I + \hat{G}(s)]^{-1} \in \mathscr{A}^{n\times n}(\sigma). \tag{24}$$

By Fact 3, (24) is equivalent to

$$\left\{\hat{Q}\left(\frac{1}{s+a}\right)[I+\hat{G}(s)]\hat{P}\left(\frac{1}{s+a}\right)\right\}^{-1} \in \hat{\mathcal{A}}^{n\times n}(\sigma).$$

Introduce now the following multiplier:

$$\hat{M}(s) \triangleq \operatorname{diag}\{\underbrace{\hat{z}(s)^m, \hat{z}(s)^m, \cdots, \hat{z}(s)^m}_{r_0}, \underbrace{\hat{z}(s)^{m-c_{r_0+1}}, \hat{z}(s)^{m-c_{r_0+2}}, \cdots, \hat{z}(s)^{m-c_\alpha}}_{\alpha - r_0}, \underbrace{1, \cdots, 1}_{n-\alpha}\} \tag{25}$$

with

$$\hat{z}(s) = \frac{s-p}{s+a} \in \hat{\mathcal{A}}(\sigma). \tag{26}$$

By (19) and (26),

$$\hat{M}(s) \in \hat{\mathcal{A}}^{n\times n}(\sigma). \tag{27}$$

Remark that

$$\left\{\hat{Q}\left(\frac{1}{s+a}\right)[I+\hat{G}(s)]\hat{P}\left(\frac{1}{s+a}\right)\right\}^{-1} = \hat{M}(s)\hat{N}(s)^{-1},$$

where

$$\hat{N}(s) \triangleq \left\{\hat{Q}\left(\frac{1}{s+a}\right)[I+\hat{G}(s)]\hat{P}\left(\frac{1}{s+a}\right)\right\}\hat{M}(s). \tag{28}$$

Clearly by (27) we are done if we can show that

$$\hat{N}(s)^{-1} \in \hat{\mathcal{A}}^{n\times n}(\sigma).$$

Because $\hat{N}(s)^{-1} = \operatorname{adj}[\hat{N}(s)] \cdot [\det \hat{N}(s)]^{-1}$, this will be established if we prove that $\hat{N}(s) \in \hat{\mathcal{A}}^{n\times n}(\sigma)$ and $\inf_{\operatorname{Re} s \geq \sigma} |\det \hat{N}(s)| > 0$ (see [1], [2]). Rewrite (25), therefore, as follows:

$$\hat{M}(s) = \hat{z}(s)^m \hat{\Delta}(s), \tag{29}$$

where

$$\hat{\Delta}(s) \triangleq \operatorname{diag}\{\underbrace{1, 1, \cdots, 1}_{r_0}, \underbrace{\hat{z}(s)^{-c_{r_0+1}}, \hat{z}(s)^{-c_{r_0+2}}, \cdots, \hat{z}(s)^{-c_\alpha}}_{\alpha - r_0}, \underbrace{\hat{z}(s)^{-m}, \hat{z}(s)^{-m}, \cdots, \hat{z}(s)^{-m}}_{n-\alpha}\}. \tag{30}$$

By (28), (13), (29), (30), (26), (14), (15), (17) and (20), we obtain

$$\hat{N}(s) = \hat{N}_1(s) + \hat{N}_2(s), \tag{31}$$

where

(a)

$$\hat{N}_1(s) = \hat{D}_1(s) \oplus \hat{D}_2(s) \tag{32}$$

with

$$\hat{D}_1(s) = \operatorname{diag}\Big\{\underbrace{\hat{a}_1\Big(\frac{1}{s+a}\Big), \hat{a}_2\Big(\frac{1}{s+a}\Big), \cdots, \hat{a}_{r_0}\Big(\frac{1}{s+a}\Big)}_{r_0},$$

$$\underbrace{\hat{b}_{r_0+1}\Big(\frac{1}{s+a}\Big), \hat{b}_{r_0+2}\Big(\frac{1}{s+a}\Big), \cdots, \hat{b}_{\alpha}\Big(\frac{1}{s+a}\Big)\Big\}}_{\alpha - r_0}, \tag{33}$$

$$\hat{D}_2(s) = \operatorname{diag}\Big\{\underbrace{\hat{b}_{\alpha+1}\Big(\frac{1}{s+a}\Big), \hat{b}_{\alpha+2}\Big(\frac{1}{s+a}\Big), \cdots, \hat{b}_{\beta}\Big(\frac{1}{s+a}\Big)}_{\beta - \alpha},$$

$$\underbrace{\hat{b}_{\beta+1}\Big(\frac{1}{s+a}\Big)\hat{z}(s)^{c_{\beta+1}-m}, \hat{b}_{\beta+2}\Big(\frac{1}{s+a}\Big)\hat{z}(s)^{c_{\beta+2}-m}, \cdots, \hat{b}_{r}\Big(\frac{1}{s+a}\Big)\hat{z}(s)^{c_r-m}}_{r-\beta}, \tag{34}$$

$$\underbrace{0, 0, \cdots, 0}_{n-r}\Big\};$$

and (b)

$$\hat{N}_2(s) = \hat{Q}\Big(\frac{1}{s+a}\Big)[I + \hat{G}_0(s)]\hat{P}\Big(\frac{1}{s+a}\Big)\hat{M}(s). \tag{35}$$

Immediately

$$\hat{N}(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma). \tag{36}$$

$\hat{N}_1(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ because all its elements are in $\hat{\mathscr{A}}(\sigma)$ (indeed all its nonzero elements are polynomials in $(1/(s+a))$ because there are no negative powers of $\hat{z}(s)$ by (21)) and $\hat{N}_2(s) \in \hat{\mathscr{A}}^{n\times n}(\sigma)$ by Fact 3, (13) and (27). Finally by (23) and since $\hat{P}(1/(s+a))$ and $\hat{Q}(1/(s+a))$ are unimodular,

$$\inf_{\operatorname{Re} s \geqq \sigma} \Big|\det \hat{Q}\Big(\frac{1}{s+a}\Big)[I + \hat{G}(s)]\hat{P}\Big(\frac{1}{s+a}\Big)\Big| > 0.$$

Hence, since by (25)–(26), $\det \hat{M}(s)$ has only one zero for $\operatorname{Re} s > \sigma$, i.e., at p, we obtain with (28):

$$\inf_{S\in U} |\det \hat{N}(s)| > 0, \tag{37}$$

where U is the half-plane $\operatorname{Re} s \geqq \sigma$ with a small neighborhood of p deleted.

Consider now $\det \hat{N}(p)$.

Remark that by (35), (22) and (25)–(26),

$$\hat{N}_2(s) = \begin{matrix} \\ \alpha\{ \end{matrix} \overbrace{\phantom{\begin{matrix}a\\b\end{matrix}}}^{\alpha} \begin{matrix} \\ \end{matrix} \begin{bmatrix} \hat{K}_{11}(s) & \hat{L}_{12}(s) \\ \hline \hat{K}_{21}(s) & \hat{L}_{22}(s) \end{bmatrix} \tag{38}$$

with

$$\hat{K}_{11}(p) = 0, \tag{39}$$

$$\hat{K}_{21}(p) = 0. \tag{40}$$

Thus by (31), (32), (38)–(40), $\det \hat{N}(p) = \det \hat{D}_1(p) \det [L_{22}(p) + \hat{D}_2(p)]$ with, by (33), (16) and (18),

$$\det \hat{D}_1(p) \neq 0, \tag{41}$$

and by (34), (18), (26) and (21),

$$\begin{aligned} \det &[\hat{L}_{22}(p) + \hat{D}_2(p)] \\ &= \det \{\hat{L}_{22}(p) + \operatorname{diag} [\hat{b}_{\alpha+1}(1/(p + a)), \cdots, \hat{b}_{\beta}(1/(p + a)), 0, \cdots, 0]\}, \end{aligned} \tag{42}$$

which is nonzero by (C). Hence

$$\det \hat{N}(p) \neq 0. \tag{43}$$

Since $\hat{N}(s)$ is continuous in $\operatorname{Re} s \geqq \sigma$, (36), (37) and (43) imply $[\hat{N}(s)]^{-1} \in \hat{\mathscr{A}}^{n \times n}(\sigma)$.

Necessity. $\hat{H} \in \hat{\mathscr{A}}^{n \times n}(\sigma)$ by assumption. Relation (23) follows immediately by [6].

To establish (C) we use contradiction. So by (42) suppose that $\det [\hat{L}_{22}(p) + \hat{D}_2(p)] = 0$. We are going to show that, for some input $u \in L_n^{2\sigma}[0, \infty)$ (i.e., $u(t)\, e^{-\sigma t} \in L_n^2[0, \infty)$), the system defined by (1)–(2) has an error e and thus also an output $y = u - e$ not in $L_n^{2\sigma}[0, \infty)$. This is a contradiction because $u \in L_n^{2\sigma}[0, \infty)$ and $H \in \mathscr{A}^{n \times n}(\sigma)$ imply $y = H * u \in L_n^{2\sigma}[0, \infty)$ (see [1], [2]).

The Laplace transforms of e and u are related by

$$[I + \hat{G}(s)]\hat{e}(s) = \hat{u}(s). \tag{44}$$

Multiply (44) on the left by $\hat{Q}((1/s + a))$ and define the n-vectors $\bar{e}(s)$ and $\bar{u}(s)$ by

$$\hat{P}\left(\frac{1}{s + a}\right) \hat{M}(s)\bar{e}(s) = \hat{e}(s), \tag{45}$$

$$\hat{Q}\left(\frac{1}{s + a}\right) \hat{u}(s) = \bar{u}(s). \tag{46}$$

By (44)–(46) and (28) obtain

$$\hat{N}(s)\bar{e}(s) = \bar{u}(s). \tag{47}$$

Because $\det [\hat{L}_{22}(p) + \hat{D}_2(p)] = 0$ we can pick a nonzero vector $\eta \in \mathbb{C}^{n-\alpha}$ in the null space of $[\hat{L}_{22}(p) + \hat{D}_2(p)]$; hence,

$$[\hat{L}_{22}(p) + \hat{D}_2(p)]\eta = 0. \tag{48}$$

Now choose the vector $\xi \in \mathbb{C}^{\alpha}$ such that

$$\xi \triangleq -[\hat{D}_1(p)]^{-1} \hat{L}_{12}(p)\eta, \tag{49}$$

which is well-defined because of (41) and the fact that all elements of $\hat{L}_{12}$ are in $\hat{\mathscr{A}}(\sigma)$.

Hence with

$$\bar{e}(s) = \frac{1}{s-p}\begin{pmatrix}\xi\\ \eta\end{pmatrix}, \tag{50}$$

and

$$\bar{u}(s) = \begin{pmatrix}\bar{u}_1(s)\\ \bar{u}_2(s)\end{pmatrix}\begin{matrix}\}\alpha\\ \}n-\alpha\end{matrix}, \tag{51}$$

and (47), (31), (32), (38), we obtain

$$\bar{u}_1(s) = \{[\hat{D}_1(s) + \hat{K}_{11}(s)]\xi + \hat{L}_{12}(s)\eta\}/(s-p), \tag{52}$$

$$\bar{u}_2(s) = \{\hat{K}_{21}(s)\xi + [\hat{D}_2(s) + \hat{L}_{22}(s)]\eta\}/(s-p). \tag{53}$$

All the components of the numerators of (52) and (53) are in $\hat{\mathscr{A}}(\sigma)$; by virtue of (39)–(40) and (48)–(49) they have at least a first order zero at p. Therefore $\bar{u}_1(s)$ and $\bar{u}_2(s)$ are well-behaved and bounded at $s = p$. Thus $\bar{u}(s)$ is analytic for $\operatorname{Re} s > \sigma$, bounded on $\operatorname{Re} s \geqq \sigma$ and as $|\omega| \to \infty$:

$$|\bar{u}(\operatorname{Re} s + j\omega)| \quad \text{is at most } O\left(\frac{1}{\omega}\right) \text{ for any fixed } \quad \operatorname{Re} s \geqq \sigma.$$

It follows therefore that the components of $\bar{u}(s)$ are the Laplace transforms of elements of $L^{2\sigma}[0, \infty)$ (see [14]). From Fact 3 and (46) we conclude that the same is true for the components of $\hat{u}(s)$; hence,

$$u \in L_n^{2\sigma}[0, \infty). \tag{54}$$

Finally by (45), (50), (25)–(26) and since $\eta \neq 0$ and $\hat{P}(1/(s+a))$ is unimodular, there exists at least one component of $\hat{e}(s)$ which has a nonzero residue at p.

Thus

$$e \notin L_n^{2\sigma}[0, \infty) \tag{55}$$

and by (54) and (55) we have established a contradiction.

Remark 1. The theorem above describes in detail what happens when K_-, K_0, K_+ are nonempty. When one or more of these sets are empty the theorem still holds provided condition (C) and the multiplier $\hat{M}(s)$ are modified in a straightforward fashion.

Remark 2. In case there are l poles at $p_1, p_2, \cdots, p_l$ of order $m_1, m_2, \cdots, m_l$ with real part larger than σ, one uses a product of multipliers like $\hat{M}(s)$, one for each pole. Condition (C) is used only to check that $\det \hat{N}(s)$ does not vanish at $s = p$. Therefore for the more general case an approximate condition (C) is required at each pole.

Remark 3. In paper [15] these techniques have been applied in a straightforward manner for the discrete-time case, thus providing a generalization to the work of Desoer, Wu and Lam [8]–[10].

Remark 4. In Theorem 3, we have used without comment the fact that the algebra $\mathcal{A}^{n\times n}(\sigma)$, which is usually defined over the field $\mathbb{R}$, extends naturally to an algebra defined over the field $\mathbb{C}$.

REFERENCES

[1] C. A. DESOER AND M. Y. WU, *Stability of linear time-invariant systems*, IEEE Trans. Circuit Theory, CT-15 (1968), pp. 245–250.

[2] ———, *Stability of multiple-loop feedback linear time-invariant systems*, J. Math. Anal. Appl., 23 (1968), pp. 121–130.

[3] R. A. BAKER AND D. J. VAKHARIA, *Input–output stability of linear time-invariant systems*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 316–319.

[4] R. E. NASBURG AND R. A. BAKER, *Stability of linear time-invariant distributed parameter single-loop feedback systems*, Ibid., AC-17 (1972), pp. 567–569.

[5] M. VIDYASAGAR, *Input–output stability of a broad class of linear time-invariant systems*, this Journal, 10 (1972), pp. 203–209.

[6] C. A. DESOER AND M. VIDYASAGAR, *General necessary conditions for input–output stability*, IEEE Proc., 59 (1971), pp. 1255–1256.

[7] C. A. DESOER AND F. L. LAM, *On the input–output properties of linear time-invariant systems*, IEEE Trans. Circuit Theory, CT-19 (1972), pp. 60–62.

[8] C. A. DESOER AND M. Y. WU, *Input–output properties of multiple-input, multiple-output discrete systems, Part I*, J. Franklin Inst., 290 (1970), no. 1, pp. 11–24.

[9] ———, *Input–output properties of multiple-input, multiple-output nonlinear discrete systems, Part II*, Ibid., 290 (1970), no. 2, pp. 85–101.

[10] C. A. DESOER AND F. L. LAM, *Stability of linear time-invariant discrete systems*, IEEE Proc., 58 (1970), pp. 1841–1843.

[11] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, 1941, Chap. 4, § 1.

[12] F. R. GANTMACHER, *Matrix Theory*, vol. I, Chelsea, New York, 1959 (pp. 130–145, esp.).

[13] K. HOFFMAN AND R. KUNZE, *Linear Algebra*, Prentice-Hall, Englewood Cliffs, N. J., 1961 (pp. 108–128, esp.).

[14] R. E. A. C. PALEY AND N. WIENER, *The Fourier Transform in the Complex Domain*, Colloquium Publications, vol. 19, American Mathematical Society, Providence, R.I., 1934 (Theorem V, p. 8, esp.).

[15] F. M. CALLIER AND C. A. DESOER, *Discrete-time convolution feedback systems*, Internat. J. Control, 15 (1972).

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

ASYMPTOTIC BEHAVIOR OF THE VALUE OF DIFFERENTIAL GAMES*

RONALD J. STERN†

Abstract. In this paper the approach to differential game theory formulated by A. Friedman [3] is employed. The major definitions and theorems used are summarized in the preliminary section of the paper.

We shall consider zero-sum two player games of fixed duration on an interval $[t_0, T]$. The goal is to derive a sufficient condition for the value of the game, $V(T)$, to approach an asymptotic limit as the duration becomes infinite. The motivation of the proof of the result obtained is Friedman's derivation of the Isaacs equation from the principle of optimality.

1. Introduction. The approach to differential game theory used here will be that of Friedman [3]. The object of this paper is to determine a sufficient condition for $V(T)$, the value as a function of duration, to approach an asymptotic limit as $T \to \infty$. The technique will resemble the method used by Friedman [3] to establish the Isaacs equation from the principle of optimality.

2. Preliminaries. Consider a system of m ordinary differential equations:

$$\dot{x} = f(t, x(t), y(t), z(t)), \qquad t_0 \leqq t \leqq T, \tag{2.1}$$

with initial condition

$$x(t_0) = x_0$$

and a payoff

$$P_T(y, z) = g(T, x(T)) + \int_{t_0}^{T} h(t, x(t), y(t), z(t))\, dt. \tag{2.2}$$

Let Y and Z be compact subsets of the Euclidean spaces R^n and R^q respectively. The controls $y(t)$ and $z(t)$ are measurable functions taking values almost everywhere in Y and Z respectively.

We shall assume:

(a) g is continuous on $[t_0, \infty) \times R^m$.

(b) h is continuous on $[t_0, \infty) \times R^m \times Y \times Z$.

The following assumptions are made concerning the dynamics:

(c) $f(t, x, y, z)$ is continuous on $[t_0, \infty) \times R^m \times Y \times Z$.

(d) There is a function $k(t)$ which belongs to $L^1(t_0, T)$ for every $T > t_0$ such that

$$|f(t, x, y, z)| \leqq k(t)(1 + |x|)$$

for all $(t, x, y, z) \in [t_0, T] \times R^m \times Y \times Z$.

* Received by the editors September 28, 1971.

† Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois. Now at Department of Business Administration, The University of Illinois, Urbana, Illinois 61801. This research was part of the author's doctoral dissertation in Applied Mathematics at Northwestern University.

(e) For each $R > 0$ there is a function $k_R(t)$ which belongs to $L^1(t_0, T)$ for every $T > t_0$ such that

$$|f(t, x, y, z) - f(t, \bar{x}, y, z)| \leqq k_R(t)|x - \bar{x}|$$

for all $t \in [t_0, T]$, $y \in Y$, $z \in Z$ and $|x| < R$, $|\bar{x}| < R$.

Assumptions (c), (d) and (e) imply that for each interval $[t_0, T]$ and pair of controls $y(t)$, $z(t)$ played on that interval, one obtains a unique trajectory $x(t)$. Hence $P_T(y, z)$ is defined for each pair of controls on each interval $[t_0, T]$.

Let n be any positive integer, and $\delta = (T - t_0)/n$. Let

$$I_j = (t_{j-1}, t_j) \quad \text{for} \quad t_j = t_0 + j\delta, \qquad 1 \leqq j \leqq n.$$

Define Y_j and Z_j to be the classes of measurable functions on I_j which almost everywhere take values in Y and Z respectively.

Let $\Gamma^{\delta,j}$ be any map of $Z_1 \times Y_1 \times Z_2 \times Y_2 \times \cdots \times Y_{j-1} \times Z_j$ into Y_j. We then call the n-tuple

$$\Gamma^\delta = (\Gamma^{\delta,1}, \cdots, \Gamma^{\delta,n})$$

an *upper-δ-strategy* for y. Similarly, we define an upper δ-strategy for z, Δ^δ, whose components $\Delta^{\delta,j}$ are maps from $Y_1 \times Z_1 \times Y_2 \times Z_2 \times \cdots \times Z_{j-1} \times Y_j$ into Z_j.

For $(2 \leqq j \leqq n)$ let $\Gamma_{\delta,j}$ be any map from $Y_1 \times Z_1 \times Y_2 \times Z_2 \times \cdots \times Y_{j-1} \times Z_{j-1}$ into Y_j, and let $\Gamma_{\delta,1}$ be any element of Y_1. We then call the n-tuple

$$\Gamma_\delta = (\Gamma_{\delta,1}, \cdots, \Gamma_{\delta,n})$$

a *lower δ-strategy* for y. Analogously one defines a lower δ-strategy Δ_δ for z.

A *constant upper (lower) δ-strategy* for y is an n-tuple as above whose components $\Gamma^{\delta,j}(\Gamma_{\delta,j})$ have only one element in Y_j for range. Similarly one defines constant δ-strategies for z.

Given a pair $(\Delta_\delta, \Gamma^\delta)$ we uniquely obtain control functions $(z_\delta(t), y^\delta(t))$, and a trajectory $x^\delta(t)$. (z_δ, y^δ) is called the *outcome* of $(\Delta_\delta, \Gamma^\delta)$.

The goal of the player y is to maximize the payoff, while the player z seeks to minimize it.

The scheme where y chooses δ-strategies Γ^δ and z chooses δ-strategies Δ_δ is called the *upper δ-game* $G^\delta(T)$. The *lower δ-game* $G_\delta(T)$ is the plan whereby y chooses δ-strategies Γ_δ and z chooses δ-strategies Δ^δ. Fixing $T \geqq t_0$ and letting $n = 1, 2, 3, \cdots$ we obtain the pair of sequences

$$G(T) = \{\{G^\delta(T)\}, \{G_\delta(T)\}\},$$

which is the *differential game* associated with (2.1) and (2.2).

The *upper δ-value* of the game played on $[t_0, T]$ is defined to be

$$V^\delta(T) = \inf_{\Delta_{\delta,1}} \sup_{\Gamma^{\delta,1}} \inf_{\Delta_{\delta,2}} \sup_{\Gamma^{\delta,2}} \cdots \inf_{\Delta_{\delta,n}} \sup_{\Gamma^{\delta,n}} P_T[\Delta_\delta, \Gamma^\delta],$$

and the *lower δ-value* of the game played on $[t_0, T]$ is

$$V_\delta(T) = \sup_{\Gamma_{\delta,1}} \inf_{\Delta^{\delta,1}} \sup_{\Gamma_{\delta,2}} \inf_{\Delta^{\delta,2}} \cdots \sup_{\Gamma_{\delta,n}} \inf_{\Delta^{\delta,n}} P_T[\Gamma_\delta, \Delta^\delta].$$

We say the differential game has *value* $V(T)$ if the limits $\lim_{\delta \to 0} V_\delta(T)$ and $\lim_{\delta \to 0} V^\delta(T)$ exist and are equal.

Before stating those results from differential game theory which we need, we shall state the following assumptions:

(f) $h(t, x, y, z) = h^1(t, x, y) + h^2(t, x, z)$.

(g) $f(t, x, y, z) = f^1(t, x, y) + f^2(t, x, z)$.

The following results are proven in [3]:

R_1. If (a)–(e) hold, then, for each $T \geqq t_0$,

$$V^\delta(T) = \inf_{\Delta_\delta} \sup_{\Gamma^\delta} P_T[\Delta_\delta, \Gamma^\delta] = \sup_{\Gamma^\delta} \inf_{\Delta_\delta} P_T[\Delta_\delta, \Gamma^\delta],$$

$$V_\delta(T) = \sup_{\Gamma_\delta} \inf_{\Delta^\delta} P_T[\Gamma_\delta, \Delta^\delta] = \inf_{\Delta^\delta} \sup_{\Gamma_\delta} P_T[\Gamma_\delta, \Delta^\delta].$$

R_2. Let (a)–(g) hold. Then the differential game has value $V(T)$ for each $T \geqq t_0$.

Instead of the partition corresponding to $\delta = (T - t_0)/n$, consider an arbitrary partition π with mesh $|\pi|$. It is clear how one goes about defining π-strategies Γ^π, Δ_π, Γ_π and Δ^π, and the upper and lower π-values $V^\pi(T)$ and $V_\pi(T)$.

R_3. Let (a)–(g) hold. Then

$$\lim_{|\pi| \to 0} V^\pi(T) = \lim_{|\pi| \to 0} V_\pi(T) = V(T) \quad \text{for each} \quad T \geqq t_0.$$

(This means the following: given $\varepsilon > 0$ there exists $\delta > 0$ such that $|V^\pi(T) - V(T)| < \varepsilon$ and $|V_\pi(T) - V(T)| < \varepsilon$ for any partition π of $[t_0, T]$ for which $|\pi| < \delta$.)

3. Asymptotic behavior of $V(T)$. The following additional assumptions are needed:

(h) $g(t, x)$ is continuously differentiable in (t, x) in $[t_0, \infty) \times R^m$.

(i) $f(t, x, y, z)$ and $h(t, x, y, z)$ are uniformly Lipschitz continuous in (t, x) on bounded subsets of $[t_0, \infty) \times R^m \times Y \times Z$.

THEOREM 3.1. *Let* (a)–(i) *hold. Then* $V(T)$ *is uniformly Lipschitz continuous on compact subsets of* (t_0, ∞).

The above is an analogue of Theorem 2.6.3 in [3]. Theorem 3.1 implies that $V(T)$ is absolutely continuous on compact subsets of (t_0, ∞). Hence $V'(T)$ exists almost everywhere on (t_0, ∞), and for any T in (t_0, ∞) we have

$$V(T) = \int_{t_0}^{T} V'(s)\, ds. \tag{3.1}$$

DEFINITION. The differential game associated with (2.1), (2.2) is *stable* at ∞ if there exists a real number L such that $V(T) \to L$ as $T \to \infty$.

Denote by $p(t)$ a continuous function on $[t_0, \infty)$ such that for each $t \in [t_0, \infty)$ we have

$$|x(t)| \leqq p(t) \tag{3.2}$$

for all possible trajectories $x(t)$ resulting from control functions $y(t)$ and $z(t)$. Such a function $p(t)$ can always be found.

THEOREM 3.2. *Let* (a)–(i) *hold. Then for each* $T \in (t_0, \infty)$, *where* $V(T)$ *is differentiable, the following holds*:

$$V'(T) \leqq \max_{|x| \leqq p(T)} \max_{y \in Y} \min_{z \in Z} \{\nabla_x g(T, x) \cdot f(T, x, y, z) + h(T, x, y, z) + \nabla_t g(T, x)\}, \tag{3.3}$$

$$(3.4)\quad V'(T) \geqq \min_{|x| \leqq p(T)} \max_{y \in Y} \min_{z \in Z} \{\nabla_x g(T, x) \cdot f(T, x, y, z) + h(T, x, y, z) + \nabla_t g(T, x)\}.$$

Before giving a proof we can state the following.

COROLLARY 3.3. *Let* (a)–(i) *hold. If the right sides of* (3.3) *and* (3.4) *are absolutely integrable on* (t_0, ∞), *the differential game associated with* (2.1)–(2.2) *is stable.*

Proof of Theorem 3.2. Let $T \geqq t_0$ be an instant where $V'(T)$ exists. Let $\overline{T} > T$. For a given positive integer n let

$$\delta = \frac{T - t_0}{n}, \quad \bar{\delta} = \frac{\overline{T} - T}{n}.$$

To save on notation we denote $V^{\delta}(T)$ by $V^n(T)$. Also, let $V^n(\overline{T})$ denote the upper π-value of the differential game played on $[t_0, \overline{T}]$, where π is the partition of $[t_0, \overline{T}]$ which is indicated above. Thus we have

$$(3.5)\quad V^n(T) = \inf_{\Delta_{\bar{\delta}}} \sup_{\Gamma^{\delta}} \left\{ g(T, x^{\delta}(T)) + \int_{t_0}^{T} h(t, x^{\delta}(t), y^{\delta}(t), z_{\bar{\delta}}(t))\, dt \right\}$$

and

$$(3.6)\quad \begin{aligned} V^n(\overline{T}) = \inf_{\Delta_{\bar{\delta}}} \sup_{\Gamma^{\delta}} \Bigg\{ & \int_{t_0}^{T} h(t, x^{\delta}(t), y^{\delta}(t), z_{\bar{\delta}}(t))\, dt \\ & + \inf_{\Delta_{\bar{\delta}}} \sup_{\Gamma^{\delta}} \{ g(\overline{T}, x^{\bar{\delta}}(\overline{T})) + \int_{T}^{\overline{T}} h(t, x^{\bar{\delta}}(t), y^{\bar{\delta}}(t), z_{\bar{\delta}}(t))\, dt \} \Bigg\}. \end{aligned}$$

Expression (3.6) follows from the fact that

$$\inf_{\Delta_{\delta,1}} \sup_{\Gamma^{\delta,1}} \inf_{\Delta_{\delta,2}} \sup_{\Gamma^{\delta,2}} \cdots \inf_{\Delta_{\delta,n}} \sup_{\Gamma^{\delta,n}} \inf_{\Delta_{\bar{\delta},1}} \sup_{\Gamma^{\bar{\delta},1}} \inf_{\Delta_{\bar{\delta},2}} \sup_{\Gamma^{\bar{\delta},2}} \cdots \inf_{\Delta_{\bar{\delta},n}} \sup_{\Gamma^{\bar{\delta},n}} = \inf_{\Delta_{\delta}} \sup_{\Gamma^{\delta}} \inf_{\Delta_{\bar{\delta}}} \sup_{\Gamma^{\bar{\delta}}}$$

which follows from R_1. Expression (3.6) resembles the principle of optimality in [3] which in turn resembles the Bellman principle commonly applied in one player decision problems [1].

Define $X(T) = \{x \in R^m : |x| \leqq p(T)\}$. From (3.5)–(3.6) we obtain

$$(3.7)\quad \begin{aligned} V^n(\overline{T}) - V^n(T) \leqq \max_{x^{\delta}(T) \in X(T)} \inf_{\Delta_{\bar{\delta}}} \sup_{\Gamma^{\delta}} \Bigg\{ & g(\overline{T}, x^{\bar{\delta}}(\overline{T})) - g(T, x^{\delta}(T)) \\ & + \int_{T}^{\overline{T}} h(t, x^{\bar{\delta}}(t), y^{\bar{\delta}}(t), z_{\bar{\delta}}(t))\, dt \Bigg\} \end{aligned}$$

and

$$(3.8)\quad \begin{aligned} V^n(\overline{T}) - V^n(T) \geqq \min_{x^{\delta}(T) \in X(T)} \inf_{\Delta_{\bar{\delta}}} \sup_{\Gamma^{\bar{\delta}}} \Bigg\{ & g(\overline{T}, x^{\bar{\delta}}(\overline{T})) - g(T, x^{\bar{\delta}}(T)) \\ & + \int_{T}^{\overline{T}} h(t, x^{\bar{\delta}}(t), y^{\bar{\delta}}(t), z_{\bar{\delta}}(t))\, dt \Bigg\}. \end{aligned}$$

We shall prove the following holds for each n and any $x^{\delta}(T) \in X(T)$:

$$(3.9)\quad \begin{aligned} \lim_{T \downarrow T} \Bigg[\inf_{\Delta_{\bar{\delta}}} \sup_{\Gamma^{\bar{\delta}}} \Bigg\{ & g(\overline{T}, x^{\bar{\delta}}(\overline{T})) - g(T, x^{\bar{\delta}}(T)) \\ & + \int_{T}^{\overline{T}} h(t, x^{\bar{\delta}}(t), y^{\bar{\delta}}(t), z_{\bar{\delta}}(t))\, dt \Bigg\} \frac{1}{(\overline{T} - T)} \Bigg] \end{aligned}$$

exists and is equal to

$$\max_{y\in Y}\min_{z\in Z}\{\nabla_x g(T, x^{\bar\delta}(T))\cdot f(T, x^{\bar\delta}(T), y, z) + h(T, x^{\bar\delta}(T), y, z) + \nabla_t g(T, x^{\bar\delta}(T))\}.$$

Since (3.9) holds for any n, R_3, (3.7), (3.8), and the compactness of $X(T)$ easily imply the assertion of Theorem 3.2.

To establish (3.9) suppose it were not true. In particular, let $x^{\bar\delta}(T)\in X(T)$ be a point where (3.9) failed to hold. Then there exists $\varepsilon > 0$ with the following property: For any $\eta > 0$, one can find $\overline{T} < T + \eta$ such that for any given $\Delta_{\bar\delta}$ there is a $\Gamma^{\bar\delta}$ for which

$$\left[g(\overline{T}, x^{\bar\delta}(\overline{T})) - g(T, x^{\bar\delta}(T)) + \int_T^{\overline{T}} h(t, x^{\bar\delta}(t), y^{\bar\delta}(t), z_{\bar\delta}(t))\,dt\right]\frac{1}{(\overline{T}-T)} - \max_{y\in Y}\min_{z\in Z}\{\nabla_x g(T, x^{\bar\delta}(T))\cdot f(T, x^{\bar\delta}(T), y, z) + h(T, x^{\bar\delta}(T), y, z) + \nabla_t g(T, x^{\bar\delta}(T))\} > \varepsilon, \tag{3.10}$$

or if $\Gamma^{\bar\delta}$ is given we could find $\Delta_{\bar\delta}$ such that the quantity on the left side of (3.10) is $< -\varepsilon$. We shall refute only the first alternative, since the case for the second is similar.

Let y^* and z^* be the points where the max min in (3.9) is attained. Define $X(x^{\bar\delta}(T), \overline{T})$ to be the set of points $x\in R^m$ such that $x = x^{\bar\delta}(t)$, where $x^{\bar\delta}$ is any trajectory emanating from $x^{\bar\delta}(T)$, and t is any instant in $[T, \overline{T}]$. η may be taken so small that the following holds:

$$\max_{(t,x)\in[T,\overline{T}]\times X(x^{\bar\delta}(T),\overline{T})} |\nabla_x g(t, x)\cdot f(t, x, y, z^*) - \nabla_x g(T, x^{\bar\delta}(T))\cdot f(T, x^{\bar\delta}(T), y, z^*) + \nabla_t g(t, x) - \nabla_t g(T, x^{\bar\delta}(T)) + h(t, x, y, z^*) - h(T, x^{\bar\delta}(T), y, z^*)| < \frac{\varepsilon}{2} \tag{3.11}$$

for all $y\in Y$.

Next, we rewrite expression (3.10) as follows:

$$\int_T^{\overline{T}} [\nabla_x g(t, x^{\bar\delta}(t))\cdot f(t, x^{\bar\delta}(t), y^{\bar\delta}(t), z_{\bar\delta}(t)) + \nabla_t g(t, x^{\bar\delta}(t)) + h(t, x^{\bar\delta}(t), y^{\bar\delta}(t), z_{\bar\delta}(t))]\,dt - [\nabla_x g(T, x^{\bar\delta}(T))\cdot f(T, x^{\bar\delta}(T), y^*, z^*) + h(T, x^{\bar\delta}(T), y^*, z^*) + \nabla_t g(T, x^{\bar\delta}(T))](\overline{T}-T) > \varepsilon(\overline{T}-T). \tag{3.12}$$

If we let $\Delta_{\bar\delta}$ be the constant lower-δ strategy for z^*, (3.12) then gives the following:

$$\int_T^{\overline{T}} [\nabla_x g(t, x^{\bar\delta}(t))\cdot f(t, x^{\bar\delta}(t), y^{\bar\delta}(t), z^*) + \nabla_t g(t, x^{\bar\delta}(t)) + h(t, x^{\bar\delta}(t), y^{\bar\delta}(t), z^*)]\,dt - [\nabla_x g(T, x^{\bar\delta}(T))\cdot f(T, x^{\bar\delta}(T), y^*, z^*) + h(T, x^{\bar\delta}(T), y^*, z^*) + \nabla_t g(T, x^{\bar\delta}(T))](\overline{T}-T) > \varepsilon(\overline{T}-T). \tag{3.13}$$

This contradicts (3.11). Thus (3.9) holds, and the proof is complete.

Remark 3.4. Let (a)–(i) hold. Assume that $\lim_{T\to\infty} p(T) = 0$ and suppose $\lim_{T\to\infty}\max_{|x|\leq p(T)} g(T, x)$ exists. Then stability of the differential game associated

with (2.1)–(2.2) is equivalent to the stability of the differential game associated with (2.1) and the payoff

$$\bar{P}_T(y, z) = \int_{t_0}^{T} h(t, x(t), y(t), z(t))\, dt.$$

Proof. For each δ and T we have

$$|\inf_{\Delta_\delta} \sup_{\Gamma^\delta} P_T(y, z) - \inf_{\Delta_\delta} \sup_{\Gamma^\delta} \bar{P}_T(y, z)| \leqq |\max_{|x| \leqq p(T)} g(T, x)|. \tag{3.14}$$

Letting $\bar{V}(T)$ denote the value of the game with payoff $\bar{P}_T$ we have

$$|V(T) - \bar{V}(T)| \leqq |\max_{|x| \leqq p(T)} g(T, x)| \tag{3.15}$$

from which the required equivalency directly follows.

4. Examples. We shall consider a certain linear-quadratic differential game. The dynamics are given by

$$\begin{aligned} \dot{x} &= A(t)x(t) + B(t)y(t) + C(t)z(t), \\ x(t_0) &= x_0. \end{aligned} \tag{4.1}$$

Here $A(t)$, $B(t)$, and $C(t)$ are continuous $m \times m$, $n \times m$ and $q \times m$ matrices, respectively, on $[t_0, \infty)$. The payoff is given by

$$P_T(y, z) = \alpha(T)|x(T)|^2 + \int_{t_0}^{T} r(t)|x(t)|^2\, dt + \int_{t_0}^{T} |z(t)|^2\, dt - \int_{t_0}^{T} |y(t)|^2\, dt. \tag{4.2}$$

$\alpha(T)$ is assumed to be positive and continuously differentiable on $[t_0, \infty)$. $r(t)$ is positive and continuous on $[t_0, \infty)$.

Case 1. We shall make the following additional assumptions for this case:

(i) There exists $T' \geqq t_0$ such that $r(T) \geqq |\alpha'(T)|$ when $T \geqq T'$.

(ii) Y and Z have interior with respect to R^n and R^q, respectively, and $0 \in \text{int } Y$, $0 \in \text{int } Z$.

(iii) There exists $T'' \geqq t_0$ such that if $T \geqq T''$, then $\alpha(T)B^t(T)x \in Y$ and $-\alpha(T)C^t(T)x \in Z$ for each $x \in R^m$ for which $|x| \leqq p(T)$, where we take

$$p(T) = |S(T, t_0)x_0| + \int_{t_0}^{T} |S(T, \tau)|[|B(\tau)| \cdot \sup_{y \in Y} |y| + |C(\tau)| \cdot \sup_{z \in Z} |z|]\, d\tau,$$

$S(T, \tau)$ being the fundamental solution of $\dot{x} + A(t)x = 0$.

(iv) There exists $T''' \geqq t_0$ such that if $T \geqq T'''$, then the matrix $2\alpha(T)A(T) + r(T)I + \alpha(T)B(T)B^t(T) - \alpha(T)C(T)C^t(T)$ is positive definite.

For this differential game we have

$$\begin{aligned} &\nabla_x g(T, x) \cdot f(T, x, y, z) + h(T, x, y, z) \\ &\quad = 2\alpha(T)x \cdot [A(T)x + B(T)y + C(T)z] + r(T)|x|^2 + |z|^2 - |y|^2. \end{aligned} \tag{4.3}$$

Denote $\bar{T} = \max\{T', T'', T'''\}$.

From assumption (iii) it follows that for $T \geqq \bar{T}$ and $|x| \leqq p(T)$ we have

$$\begin{aligned} &\max_{y \in Y} \min_{z \in Z} \{\nabla_x g(T, x) \cdot f(T, x, y, z) + h(T, x, y, z)\} \\ &\quad = 2\alpha(T)(x \cdot A(T)x) + r(T)|x|^2 + \alpha^2(T)|B^t(T)x|^2 - \alpha^2(T)|C^t(T)x|^2. \end{aligned} \tag{4.4}$$

Equation (4.4), assumption (i) and (3.3) yield

$$V'(T) \leqq [2\alpha(T)|A(T)| + r(T) + \alpha^2(T)|B(T)|^2 + \alpha'(T)]p^2(T) \quad \text{if} \quad T \geqq \bar{T}. \tag{4.5}$$

Equation (4.4), assumption (iv) and (3.4) yield

$$V'(T) \geqq 0 \quad \text{if} \quad T \geqq \bar{T}. \tag{4.6}$$

Thus, if the right side of (4.5) is integrable, the differential game is stable. In the next case to be considered we shall impose conditions on the dynamics that result in a more readily verifiable sufficient condition for stability.

Case 2. We shall assume the following:

(i′) $A(t) = A$, where A is a matrix having the real parts of its eigenvalues strictly negative.

(ii′) $|B(t)| \to 0$ and $|C(t)| \to 0$ as $t \to \infty$.

(iii′) $\sup_{y \in Y}|y| = \sup_{z \in Z}|z|$.

(iv′) $\alpha(T)$ is bounded on $[t_0, \infty)$.

Using methods in [1, pp. 327–328], we find that there exist positive constants c_1 and c_2 such that

$$|x(t)| \leqq c_1 e^{-c_2 t} + c_1|q(t_0)|e^{-c_2 t} + c_1\left|q\left(\frac{t}{2}\right)\right|, \qquad t \geqq 2t_0, \tag{4.7}$$

where $q(t)$ is any continuous function such that $|B(t)| + |C(t)| \leqq q(t)$ and $q(t) \to 0$ monotonely as $t \to \infty$. Thus, Remark 3.4 can be applied.

Consider the modified game with payoff given by

$$\bar{P}_T(y, z) = \int_{t_0}^{T} r(t)|x(t)|^2\,dt + \int_{t_0}^{T} |z(t)|^2\,dt - \int_{t_0}^{T} |y(t)|^2\,dt. \tag{4.8}$$

It follows from assumption (iii′) that for this game we have

$$\max_{y \in Y} \min_{z \in Z} h(T, x, y, z) = r(T)|x|^2. \tag{4.9}$$

Now applying Theorem 3.2 we have

$$0 \leqq \bar{V}'(T) \leqq r(T)p^2(T), \tag{4.10}$$

where $p(T)$ is given by the right side of (4.7). Thus, the unmodified game is stable if $q(t)$ is integrable.

REFERENCES

[1] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, N.J., 1957.

[2] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[3] A. Friedman, *Differential Games*, Interscience, New York, 1971.

SIAM J. CONTROL
Vol. 10, No. 4, November 1972

A MARTINGALE APPROACH TO LINEAR RECURSIVE STATE ESTIMATION*

A. V. BALAKRISHNAN†

Abstract. In this paper we present a new approach to linear recursive estimation (Bucy–Kalman filtering) which we believe may be more elegant and satisfying than extant treatments. The main idea is to exploit the simpler problem of estimating one Martingale from another.

In this paper we present a new approach to Kalman filtering which we believe may be more elegant and satisfying than extant treatments [1], [2], [3] including the innovations approach of Kailath [3]. The main idea is to exploit the simpler problem of estimating one Martingale from another. Succinctly stated, for the system

$$x(t;\omega) = \int_0^t A(s)x(s;\omega)\,ds + \int_0^t B(s)\,dW(s;\omega),$$

$$Y(t;\omega) = \int_0^t C(s)x(s;\omega)\,ds + \int_0^t D(s)\,dW(s;\omega),$$

where $W(t;\omega)$ is a Wiener process and

$$D(s)D(s)^* > 0,$$

by letting, in the usual notation,

$$\hat{x}(t;\omega) = E[x(t;\omega)|\, Y(s;\omega), 0 \leqq s \leqq t],$$

we note

$$Y(t) - \int_0^t C(s)\hat{x}(s;\omega)\,ds = Z_0(t;\omega)$$

is a Martingale (this result is known; see [1], [3]).

But it is also true that

$$\hat{x}(t;\omega) - \int_0^t A(s)\hat{x}(s;\omega)\,ds = Z_s(t;\omega)$$

is also a Martingale.

Given the two Martingales $Z_0(t;\omega)$, $Z_s(t;\omega)$, the best mean square estimate $\hat{Z}_s(t;\omega)$ of $Z_s(t;\omega)$ from $Z_0(\sigma;\omega)$, $\sigma \leqq t$, is given by

$$\hat{Z}_s(t;\omega) = \int_0^t \lambda_{12}(s)\,dZ_0(s;\omega).$$

We give a simple proof of the equivalence of the sigma algebras generated by $Z_0(t;\omega)$ and $Y(t;\omega)$ so that $Z_s(t;\omega) = \hat{Z}_s(t;\omega)$. We use some results of E. Nelson's recent work [4] on Martingales, in a crucial way.

* *Received by the editors June 24, 1971.*

† System Science Department, School of Engineering and Applied Science, University of California, Los Angeles, California 90024. This research was supported in part by the United States Air Force, Applied Mathematics Division, under Grant AFOSR 68-1408.

For convenience, all vectors are taken as columns ($n \times 1$ matrices) and * denotes adjoint. Also $\| \cdot \|$ denotes the Euclidean norm:

$$\|A\|^2 = \operatorname{tr} AA^*.$$

2. We begin with a fundamental property of Martingales (cf. Nelson [4], Doob [5]).

LEMMA. *Let $Z_i(t;\omega)$, $i = 1, 2$, denote two Martingales with respect to the growing sigma algebra $\mathscr{F}(t)$, and let*

$$E\|Z_i(1;\omega) - Z_i(0;\omega)\|^2 < \infty, \qquad i = 1, 2 \cdots. \tag{2.1}$$

Suppose that for i, j fixed, $0 \leqq t < 1$,

$$\lim_{\Delta \to 0} \frac{1}{\Delta} E((Z_i(t+\Delta;\omega) - Z_i(t;\omega))(Z_j(t+\Delta;\omega) - Z_j(t;\omega))^* | \mathscr{F}(t)) = P_{ij}(t), \tag{2.2}$$

where the convergence of the random variable on the left is in the mean order one (L_1), and assume that $P_{ij}(t)$ is continuous in t, $0 \leqq t \leqq 1$. Then for $0 \leqq s \leqq t \leqq 1$,

$$E\left(\left(\int_s^t dZ_i(\sigma;\omega)\right)\left(\int_s^t dZ_j(\sigma;\omega)\right)^* | \mathscr{F}(s)\right) = \int_s^t P_{ij}(\sigma)\, d\sigma. \tag{2.3}$$

Proof. Let $0 \leqq a < 1$, and define

$$\Lambda_{ij}(t) = \left(\int_a^t dZ_i(s;\omega)\right)\left(\int_a^t dZ_j(s;\omega)\right)^*, \qquad a \leqq t \leqq 1.$$

Observe that for any $\Delta > 0$, $t + \Delta < 1$, we have

$$\begin{aligned} &E((\Lambda_{ij}(t+\Delta) - \Lambda_{ij}(t))|\mathscr{F}(a)) \\ &= E\left[E\left[\left(\int_t^{t+\Delta} dZ_i(s;\omega)\right)\left(\int_t^{t+\Delta} dZ_j(s;\omega)\right)^* \middle| \mathscr{F}(t)\right] \middle| \mathscr{F}(a)\right], \end{aligned}$$

where we have used the fact that

$$E[[(Z_i(t+\Delta;\omega) - Z_i(t;\omega))(Z_j(t;\omega) - Z_j(a;\omega))^*]|\mathscr{F}(t)] = 0.$$

We can now follow the argument of Nelson [4]. Let $\varepsilon > 0$ be given. Let J denote the set of points in $[a, 1]$ such that for t in J,

$$E\left(\|E(\Lambda_{ij}(t)|\mathscr{F}(a)) - \int_a^t P_{ij}(s)\, ds\|\right) \leqq \varepsilon(t - a). \tag{2.4}$$

Clearly J contains the point a. Also J is closed. Now we shall show that if t is any point for which (2.4) holds, then (2.4) will hold for $t + \delta, 0 < \delta < \Delta$, for some $\Delta > 0$. For this, we note that

$$E(\Delta_{ij}(t+\delta)|\mathscr{F}(a)) = E[\Lambda_{ij}(t+\delta) - \Lambda_{ij}(t)|\mathscr{F}(t)|\mathscr{F}(a)] + E(\Lambda_{ij}(t)|\mathscr{F}(a)),$$

and we choose Δ_1 such that for all $\delta < \Delta_1$,

$$E\|(E(\Lambda_{ij}(t+\delta) - \Lambda_{ij}(t)|\mathscr{F}(t)) - \delta P_{ij}(t))\| < (\varepsilon/2)\delta.$$

Next let us choose Δ_2 such that for all $\delta < \Delta_2$,

$$\int_t^{t+\delta} P_{ij}(s)\,ds - \delta P_{ij}(t) \ < (\varepsilon/2)\delta.$$

Choosing Δ to be the minimum of Δ_1 and Δ_2, we have, for $0 \leqq \delta \leqq \Delta$,

$$E\left|E(\Lambda_{ij}(t+\delta) - \Lambda_{ij}(t)|\mathscr{F}(a)) - \int_t^{t+\delta} P_{ij}(s)\,ds\right|$$

$$= E\left\|E\left\{E(\Lambda_{ij}(t+\delta) - \Lambda_{ij}(t)|\mathscr{F}(t)) - \int_t^{t+\delta} P_{ij}(s)\,ds\right\}|\mathscr{F}(a)\right\|$$

$$\leqq E\left\|E(\Lambda_{ij}(t+\delta) - \Lambda_{ij}(t)|\mathscr{F}(t)) - \int_t^{t+\delta} P_{ij}(s)\,ds\right\| \leqq (\varepsilon)\delta.$$

Hence (2.4) follows for $t + \delta, 0 < \delta < \Delta$. In particular this is true for $t = a$.

Suppose now that the upper bound of t such that (2.4) holds for $[a, t]$ is t_0. Then t_0 must belong to J since J is closed. But if t_0 is not equal to 1, we will have a contradiction because it can be extended by a nonzero amount. Since ε is arbitrary, it is clear that

$$E[\Lambda_{ij}(t)|\mathscr{F}(a)] = \int_a^t P_{ij}(s)\,ds,$$

which is clearly enough to prove the lemma.

COROLLARY. *Assume now that the Martingales $Z_i(s;\omega)$ are Gaussian, $0 \leqq s \leqq 1$, and that*

$$Z_i(0;\omega) = 0, \qquad i = 1, 2.$$

Assume further that (2.1) *holds, and that* (2.2) *holds for $i = j = 2$, and for $i = 1, j = 2$. Let $\beta_2(t)$ be the smallest sigma algebra generated by $Z_2(s;\omega)$, $s \leqq t$. Then*

$$E(Z_1(t;\omega)|\beta_2(t)) = \int_0^t r_{12}(s)\,dZ_2(s;\omega), \qquad 0 < t \leqq 1, \tag{2.5}$$

where $r_{12}(s)$ is defined as the limit

$$r_{12}(s) = \lim_{\varepsilon \to 0} P_{12}(s)(P_{22}(s) + \varepsilon I)^{-1} \quad a.e.\ 0 < s < 1. \tag{2.6}$$

Proof. It should be emphasized that it is *not* assumed that (2.2) holds for $i = j = 1$. First let us prove that for any Lebesgue measurable matrix function $f(\cdot)$, having the same dimension as $P_{12}(s)$, such that

$$\int_0^1 \operatorname{tr} f(s)^* f(s) P_{22}(s)\,ds < \infty, \tag{2.7}$$

we have that

$$E\left(\int_0^1 dZ_1(s;\omega)\right)\left(\int_0^1 f(s)\,dZ_2(s;\omega)\right)^* = \int_0^1 P_{12}(s) f(s)^*\,ds. \tag{2.8}$$

For this it is enough to note that for $0 < s < t < 1$,

$$E\left(\int_0^1 dZ_1(s;\omega)\right)(Z_2(t;\omega) - Z_2(s;\omega))^*$$

$$= E(Z_1(t;\omega) - Z_1(s;\omega))(Z_2(t;\omega) - Z_2(s;\omega))^* = \int_s^t P_{12}(\sigma)\,d\sigma$$

from (2.3); hence it is immediate that (2.8) will hold for simple functions, and hence by the usual limiting arguments for any $f(\cdot)$ satisfying (2.7). Next if

$$L(\varepsilon;t) = P_{12}(t)(P_{22}(t) + \varepsilon I)^{-1},$$

we have that $L(\varepsilon;t)$ satisfies (2.7), and that

$$0 \leqq E\left\|Z_1(1;\omega) - \int_0^1 L(\varepsilon;t)\,dZ_2(t;\omega)\right\|^2$$

$$= E\|Z_1(1;\omega)\|^2 - \operatorname{tr}\int_0^1 P_{12}(t)(P_{22}(t) + \varepsilon I)^{-1}(P_{22}(t) + 2\varepsilon I)$$

$$\cdot(P_{22}(t) + \varepsilon I)^{-1}P_{12}(t)^*\,dt.$$

By Fatou's lemma, it follows that r_{12} satisfies (2.7). Hence the integral

$$\int_0^t r_{12}(s)\,dZ_2(s;\omega), \qquad 0 < t < 1,$$

is well-defined. Let $f(\cdot)$ satisfy (2.7). Then

$$E\left(Z_1(t;\omega) - \int_0^t r_{12}(s)\,dZ_2(s;\omega)\right)\left(\int_0^t f(s)\,dZ_2(s;\omega)\right)^*$$

$$= \int_0^t (P_{12}(s)f(s)^* - r_{12}(s)P_{22}(s)f(s)^*)\,ds$$

But

$$r_{12}(s)P_{22}(s) = P_{12}(s)\lim_{\varepsilon\to 0}(P_{22}(s) + \varepsilon I)^{-1}P_{22}(s) = P_{12}(s),$$

which is trivially true at a point where $P_{22}(s) > 0$ and otherwise true because if $P_{22}(s)x = 0$ then $P_{12}(s)x = 0$, as can be verified from (2.2). Hence, $Z_1(t;\omega) - \int_0^t r_{12}(s)\,dZ_2(s;\omega)$ is uncorrelated with, and being Gaussian, independent of, $\int_0^t f(s)\,dZ_2(s;\omega)$ for every $f(\cdot)$ satisfying (2.7). But since the latter generate $\beta_2(t)$, (2.5) follows.

Let us consider now the linear stochastic equation

$$x(t;\omega) = \int_0^t A(s)x(s;\omega)\,ds + \int_0^t B(s)\,dW(s;\omega), \tag{2.9}$$

$$Y(t;\omega) = \int_0^t C(s)x(s;\omega)\,ds + \int_0^t D(s)\,dW(s;\omega), \tag{2.10}$$

where we shall for simplicity assume that all coefficients are continuous. Further we shall assume that

$A(s)$ is $m \times m$, $B(s)$ is $m \times n$,
$C(s)$ is $q \times m$, $D(s)$ is $q \times n$,
$D(s)D(s)^* > 0$ on $[0, 1]$.

Note that this implies that $b(s) = \sqrt{(D(s)D(s)^*)^{-1}}$ is continuous. Then as we have seen, defining

$$\hat{Y}(t;\omega) = \int_0^t b(s)\, dY(s;\omega)$$

we have

$$\hat{Y}(t;\omega) = \int_0^t b(s)C(s)x(s;\omega)\, ds + \hat{W}(t;\omega),$$

where $\hat{W}(s;\omega)$ is now a Wiener process. Let $\hat{W}(t;\omega)$ be measurable with respect to the growing sigma algebra $\mathscr{F}(t)$. Then of course $x(t;\omega)$, $Y(t;\omega)$, $\hat{Y}(t;\omega)$, $\hat{W}(t;\omega)$ are all measureable $\mathscr{F}(t)$. Let

$$\mathscr{B}_Y(t) = \text{sigma algebra generated by } Y(s;\omega), \qquad s \leqq t,$$

and similarly define $\mathscr{B}_{\hat{Y}}(t)$. Then since

$$Y(t;\omega) = \int_0^t (b(s))^{-1}\, d\hat{Y}(s;\omega),$$

we have that

$$\mathscr{B}_Y(t) = \mathscr{B}_{\hat{Y}}(t).$$

Since we have that $E(|x(t;\omega)|^2) < \infty$ we can define the conditional expectation

$$\hat{x}(t;\omega) = E(x(t;\omega)|\mathscr{B}(t)),$$

where from now on we write $\mathscr{B}(t)$ for $\mathscr{B}_Y(t) = \mathscr{B}_{\hat{Y}}(t)$.

Letting $\phi(t)$ denote a fundamental matrix solution of $\dot{x}(t) = A(t)x(t)$, observe now that

$$E(x(t;\omega)|\mathscr{B}(s)) = E\left(E\left(\phi(t)\phi(s)^{-1}x(s;\omega) + \phi(t)\int_s^t \phi(\sigma)^{-1}B(\sigma)dW(\sigma;\omega)\,|\mathscr{F}(s)\right)|\mathscr{B}(s)\right)$$

$$= \phi(t)\phi(s)^{-1}\hat{x}(s;\omega).$$

But the left side is a Martingale in s, $s < t$, for fixed t, and converges with probability one as s goes to t. But $Y(t;\omega)$ being continuous with probability one, we note that

$$\mathscr{B}(t) = \text{smallest sigma algebra containing } \mathscr{B}(s) \quad \text{for every } s < t,$$

and hence the limit must be $\hat{x}(t;\omega)$. Thus $\hat{x}(t;\omega)$ is continuous from below with probability one. Hence by a theorem and a remark of Doob [5, Theorem 2.6, p. 61, and Remark, p. 62], there is a process equivalent to $\hat{x}(t,\omega)$ which is jointly measurable in t and ω, and we may clearly redefine $\hat{x}(t;\omega)$ to be this process. And since

$$E(|\hat{x}(t;\omega)|^2) \leqq E(|x(t;\omega)|^2) < \infty,$$

it follows that the integral

$$\int_0^t A(s)\hat{x}(s;\omega)\,ds, \qquad t \leqq 1,$$

is well-defined (converges a.s.).

LEMMA. *The process*

$$\hat{x}(t;\omega) - \int_0^t A(s)\hat{x}(s;\omega)\,ds = Z_s(t;\omega), \qquad 0 \leqq t \leqq 1,$$

is a Gaussian Martingale. Moreover,

$$E(|Z_s(t;\omega)|^2) < \infty.$$

Proof. Now for $t > s$, we have

$$\begin{aligned} E(\hat{x}(t;\omega)|\mathscr{B}(s)) &= E(x(t;\omega)|\mathscr{B}(t)|\mathscr{B}(s)) \\ &= E(x(t;\omega)|\mathscr{B}(s)) \cdots . \end{aligned} \tag{2.11}$$

Hence,

$$\begin{aligned} E(Z_s(t;\omega) - Z_s(s;\omega)|\mathscr{B}(s)) &= E(x(t;\omega) - x(s;\omega)|\mathscr{B}(s)) \\ &\quad - E\left(\int_s^t A(\sigma)\hat{x}(\sigma;\omega)\,d\sigma|\mathscr{B}(s)\right) \\ &= E\left(\int_s^t B(\sigma)\,dW(\sigma;\omega)|\mathscr{B}(s)\right) \\ &\quad + E\left(\int_s^t A(\sigma)(x(\sigma;\omega) - \hat{x}(\sigma;\omega))\,d\sigma|\mathscr{B}(s)\right). \end{aligned}$$

But the first term is zero because $\mathscr{F}(s) \supset \mathscr{B}(s)$, and $\int_0^t B(\sigma)\,dW(\sigma, u)$ is a Martingale; the second term is zero, by use of (2.11).

LEMMA. *Let*

$$Z_0(t;\omega) = Y(t;\omega) - \int_0^t C(s)\hat{x}(s;\omega)\,ds.$$

Then $Z_0(t, u)$ is a Gaussian Martingale. Moreover,

$$\lim_{\Delta\to 0}\frac{1}{\Delta}E\left(\int_t^{t+\Delta} dZ_0(s;\omega)\left(\int_t^{t+\Delta} dZ_0(s;\omega)^*\right)|\mathscr{B}(t)\right) = D(t)D(t)^*. \tag{2.12}$$

Proof. We have only to note that

$$\begin{aligned} Z_0(t;\omega) &= Y(t;\omega) - \int_0^t C(s)x(s;\omega)\,ds + \int_0^t C(s)(x(s;\omega) - \hat{x}(s;\omega))\,ds \\ &= \int_0^t D(s)\,dW(s;\omega) + \int_0^t C(s)(x(s;\omega) - \hat{x}(s;\omega))\,ds \end{aligned} \tag{2.13}$$

so that

$$Z_0(t;\omega) - Z_0(s;\omega) = \int_s^t D(\sigma)\,dW(\sigma;\omega) + \int_s^t C(\sigma)(x(\sigma;\omega) - \hat{x}(\sigma;\omega))\,d\sigma,$$

and just as in the previous lemma,

$$E(Z_0(t;\omega) - Z_0(s;\omega)|\mathscr{B}(s)) = 0.$$

Let us use the notation

$$e(s;\omega) = x(s;\omega) - \hat{x}(s;\omega) \tag{2.14}$$

and observe that

$$E\left(\left\|\int_t^{t+\Delta} C(s)\, e(s;\omega)\, ds\right\|^2 |\mathscr{B}(t)\right) = O(\Delta^2)$$

since

$$E(|e(s;\omega)|^2) \leqq E(|x(s;\omega)|^2)$$

and is bounded in $0 < s < 1$. Again

$$E\left(\left\|\int_t^{t+\Delta} D(s)\, dW(s;\omega)\right\|^2 |\mathscr{B}(t)\right) = O(\Delta).$$

Hence, clearly,

$$\frac{1}{\Delta}E\left(\left(\int_t^{t+\Delta} dZ_0(s;\omega)\right)\left(\int_t^{t+\Delta} dZ_0(s;\omega)\right)^* |\mathscr{B}(t)\right) = \frac{1}{\Delta}\int_t^{t+\Delta} D(s)D(s)^*\, ds + O(\Delta^{1/2}).$$

Hence (2.12) follows.

LEMMA. *Let*

$$E(e(t;\omega)\, e(t;\omega)^*) = P(t).$$

Then

$$\lim_{\Delta\to 0}\frac{1}{\Delta}E\left(\int_t^{t+\Delta} dZ_s(\sigma;\omega)\left(\int_t^{t+\Delta} dZ_0(s;\omega)\right)^* |\mathscr{B}(t)\right) = P(t)C(t)^* + B(t)D(t)^* \quad a.e. \tag{2.15}$$

Proof. Since $Z_0(s;\omega)$ satisfies (2.2) and $Z_s(t;\omega)$ satisfies (2.1), it only remains to calculate (2.15). We have seen that

$$Z_s(t+\Delta;\omega) - Z_s(t;\omega) = \int_t^{t+\Delta} A(s)\, e(s;\omega)\, ds + \int_t^{t+\Delta} B(s)\, dW(s;\omega) - e(t+\Delta;\omega) + e(t;\omega).$$

It is immediate that

$$\left|E\left(\int_t^{t+\Delta} A(s)\, e(s;\omega)\, ds\left(\int_t^{t+\Delta} dZ_0(s;\omega)\right)^* |\mathscr{B}(t)\right)\right| = O(|\Delta|^{3/2}).$$

Next, for any $\Delta > 0$, $e(t+\Delta;\omega)$ is uncorrelated with $Y(s;\omega)$, $s \leqq t+\Delta$, and hence with $Z_0(s;\omega)$, $s \leqq t+\Delta$. It is also uncorrelated with (and hence also independent of) the random variables generating $\mathscr{B}(t)$. Hence,

$$E\left(e(t+\Delta;\omega)\left(\int_t^{t+\Delta} dZ_0(s;\omega)\right)^* |\mathscr{B}(t)\right) = E(e(t+\Delta;\omega))E\left(\int_t^{t+\Delta} dZ^0(s;\omega)\right)^* = 0.$$

Since

$$Z_0(s;\omega) = \int_0^s C(\sigma)\, e(\sigma;\omega)\, d\sigma + \int_0^s D(\sigma)\, dW(\sigma;\omega)$$

we have

$$E\left(e(t;\omega)\left(\int_t^{t+\Delta} dZ_0(s;\omega)\right)^* | \mathscr{B}(t)\right) = \int_t^{t+\Delta} E(e(t;\omega)\, e(s;\omega)^* | \mathscr{B}(t))C(s)^*\, ds$$

$$= \int_t^{t+\Delta} E(e(t;\omega)\, e(s;\omega)^*)C(s)^*\, ds$$

and

$$E\left(\int_t^{t+\Delta} B(s)\, dW(s;\omega)\left(\int_t^{t+\Delta} dZ_0(s;\omega)\right)^* | \mathscr{B}(t)\right) = \int_t^{t+\Delta} B(s)D(s)^*\, ds + O(\Delta^{3/2})$$

since

$$E\left(\int_t^{t+\Delta} B(s)\, dW(s;\omega)\left(\int_t^{t+\Delta} C(s)\, e(s;\omega)\, ds\right)^* | \mathscr{B}(t)\right) = O(\Delta^{3/2}).$$

Hence (2.15) follows by taking the limit as Δ goes to zero. Note that a.e. in (2.15) is necessary since we cannot (at this stage) assert that $e(t;\omega)$ is continuous in t. Eventually we shall see that it is (immediately following the proof of Corollary 2).

LEMMA. *Under the assumption that*

$$D(s)D(s)^* > 0, \qquad 0 \leqq s \leqq 1,$$

we can write

$$\hat{x}(t;\omega) = \int_0^t k(t;s)\, dY(s;\omega), \qquad 0 \leqq t \leqq 1, \tag{2.16}$$

where

$$\int_0^1 \int_0^t |k(t;s)|^2\, ds\, dt < \infty. \tag{2.17}$$

Proof. Let us first note that

$$\int_0^t f(s)\, dY(s;\omega) = \int_0^t (f(s)D(s) + h(s))\, dW(s;\omega),$$

where

$$h(s) \int_s^t f(\sigma)C(\sigma)\phi(\sigma)\, d\sigma\phi(s)^{-1}\, B(s).$$

Define the operator L by

$$Lf = g, \qquad g(s) = f(s)D(s) + \int_s^t f(\sigma)C(\sigma)\phi(\sigma)\, d\sigma\phi(s)^{-1}B(s).$$

Then it follows that

$$E\left(\int_0^t f(s)\,dY(s;\omega)\right)\left(\int_0^t q(s)\,dY(s;\omega)\right)^* = \int_0^t p(s)q(s)^*\,ds, \tag{2.18}$$

where

$$L^*Lf = p.$$

Here $f(\cdot)$ is an $m \times q$ matrix function, and so is $p(\cdot)$. Specifically,

$$L^*h = p, \qquad p(s) = h(s)D(s)^* + \int_0^s h(\sigma)B(\sigma)^*\phi(\sigma)^{*-1}\,d\sigma\phi(s)^*C(s)^*, \tag{2.19}$$

where $h(\cdot)$ is $m \times n$ and $p(\cdot)$ is $m \times q$.

Next,

$$\begin{aligned} &E\left(x(t;\omega)\left(\int_0^t q(s)\,dY(s;\omega)\right)^*\right) \\ &\quad = E\left(\phi(t)\int_0^t \phi(s)^{-1}B(s)\,dW(s;\omega)\right)\left(\int_0^t q(s)\,dY(s;\omega)\right)^* \\ &\quad = \int_0^t r(s)q(s)^*\,ds, \end{aligned}$$

where

$$r = L^*v \quad \text{and} \quad v(s) = \phi(t)\phi(s)^{-1}B(s), \qquad 0 \leqq s \leqq t.$$

Now

$$E\left(x(t;\omega) - \int_0^t k(t;s)\,dY(s;\omega)\right)\left(\int_0^t q(s)\,dY(s;\omega)\right)^* = \int_0^t z(s)q(s)^*\,ds, \tag{2.20}$$

where

$$L^*Lk - L^*v = z,$$

where k stands for the function $k(t;\cdot)$. Hence for (2.16) to hold it is necessary and sufficient that

$$z(s) = 0 \quad \text{or} \quad L^*Lk = L^*v. \tag{2.21}$$

But because $D(s)D(s)^*$ is positive, L^*L has a bounded inverse, so that the first part of our result that there exists a function $k(t;s)$ satisfying (2.16) and such that $\int_0^t |k(t;s)|^2\,ds < \infty$ for each $0 < t < 1$, is immediate.

We need however to show that the double integral in (2.17) makes sense and that it is finite. For this we proceed to a closer examination of (2.21). Thus we note that if $Lf = g$, we can write

$$\begin{aligned} g(s) = f(s)D(s) &- \int_0^s f(\sigma)C(\sigma)\phi(\sigma)\,d\sigma\phi(s)^{-1}B(s) \\ &+ \int_0^t f(\sigma)C(\sigma)\phi(\sigma)\,d\sigma\phi(s)^{-1}\,B(s). \end{aligned}$$

The point in doing this is that the first two terms are independent of t. Again if we denote by $\tilde{L}$ the operator yielding the first two terms:

$$\tilde{L}f = g, \qquad f(s)D(s) - \int_0^s f(\sigma)C(\sigma)\phi(\sigma)\,d\sigma\phi(s)^{-1}B(s) = g(s), \tag{2.22}$$

we have

$$L^*Lf = L^*\tilde{L}f + \left(\int_0^t f(\sigma)C(\sigma)\phi(\sigma)\,d\sigma\right)(\phi(s)^{-1}B(s)D(s)^* + R(s)\phi(s)^*C(s)^*), \tag{2.23}$$

where

$$R(s) = \int_0^s \phi(\sigma)^{-1}B(\sigma)B(\sigma)^*\phi(\sigma)^{*-1}\,d\sigma.$$

Also, the function $r(\cdot)$ in (2.20) can be expressed as

$$r(s) = \phi(t)\phi(s)^{-1}B(s)D(s)^* + \phi(t)R(s)\phi(s)^*C(s)^*. \tag{2.24}$$

Hence (2.21) can be written as

$$h(t;s) = -\int_0^t k(t;\sigma)C(\sigma)\phi(\sigma)\,d\sigma u(s) + \phi(t)u(s),$$

where

$$h(t;\cdot) = L^*\tilde{L}(k(t;\cdot)),$$
$$u(s) = \phi(s)^{-1}B(s)D(s)^* + R(s)\phi(s)^*C(s)^*.$$

We now exploit the fact that $h(t;s)$ factors into a function of time and a function of s:

$$h(t;s) = \left(-\int_0^t k(t;\sigma)C(\sigma)\phi(\sigma)\,d\sigma + \phi(t)\right)u(s).$$

But since $D(s)D(s)^*$ is assumed positive, we note that L^* has a bounded inverse in $L_2(0,1)$, and so does $\tilde{L}$; moreover if

$$L^*h = p,$$

we have

$$h(s) = h_1(s)D(s),$$
$$p(s)(D(s)D(s)^*)^{-1} = h_1(s) + \int_0^s h_1(\sigma)D(\sigma)B(\sigma)^*\phi(\sigma)^{*-1}\,d\sigma\phi(s)^*C(s)^*.$$

Since the right side is "identity plus Volterra operator", we can use a Neumann expansion to find the inverse, and similarly for $\tilde{L}$. But since each of these operations does not involve t, it is readily seen that this implies that it is possible to express $k(t;s)$ as

$$k(t;s) = k_1(t)k_2(s), \tag{2.25}$$

where $k_1(\cdot)$ is $m \times m$.

Hence substituting this into (2.21), and taking advantage of the forms in (2.23) and (2.24) we must have

$$L^*\tilde{L}k_2 = u(\cdot), \qquad u(s) = \phi(s)^{-1}B(s)D(s)^* + R(s)\phi(s)^*C(s)^*, \qquad 0 < s < 1.$$

And $k_1(t)$ must satisfy

$$\phi(t) - \int_0^t k_1(t)k_2(s)C(s)\phi(s)\,ds = k_1(t).$$

But since $\phi(t)$ is nonsingular, it follows that both $k_1(t)$ and $(I + \int_0^t k_2(s)C(s)\phi(s)\,ds)$ are nonsingular, and hence

$$k_1(t) = \phi(t)\left(I + \int_0^t k_2(s)C(s)\phi(s)\,ds\right)^{-1} \tag{2.26}$$

from which (2.17) follows. Of course we have obtained more than we sought to prove.

Now we can prove one of the main theorems [1].

THEOREM 2.7. *Under the assumption that*

$$D(s)D(s)^* > 0$$

for every $s, 0 \leqq s \leqq 1$, *we have for every* $t, 0 \leqq t \leqq 1$,

$$\mathscr{B}(t) = \textit{smallest sigma algebra generated by } \{Z_0(s;\omega), s \leqq t\}.$$

Proof. Let us recall that

$$Z_0(s;\omega) = Y(s;\omega) - \int_0^s C(\sigma)\hat{x}(\sigma;\omega)\,d\sigma. \tag{2.27}$$

Hence for any $q \times q$ matrix function $f(\cdot)$ in (appropriate-dimensional) $L_2(0,t)$-space, we have

$$\int_0^t f(s)\,dZ_0(s;\omega) = \int_0^t f(s)\,dY(s;\omega) - \int_0^t f(s)C(s)\hat{x}(s;\omega)\,ds.$$

But using (2.16), we have

$$\begin{aligned}\int_0^t f(s)C(s)\hat{x}(s;\omega)\,ds &= \int_0^t f(s)C(s)\int_0^s k(s;\sigma)\,dY(\sigma;\omega)\,ds \\ &= \int_0^t \left(\int_\sigma^t f(s)C(s)k(s;\sigma)\,ds\right) dY(\sigma;\omega).\end{aligned}$$

Introduce now the operator

$$Hf = g, \qquad g(\sigma) = f(\sigma) - \int_\sigma^t f(s)C(s)k(s;\sigma)\,ds, \qquad 0 \leqq \sigma \leqq t,$$

which maps $L_2(0,t)$ into itself. More importantly, in view of (2.17) it differs from the identity operator by a Volterra operator with a square integrable kernel. Hence H has a bounded inverse.

Hence for any $g(\cdot)$ in $L_2(0, t)$,

$$\int_0^t g(s)\, dY(s;\omega) = \int_0^t f(s)\, dZ_0(s;\omega), \qquad f = H^{-1}g.$$

Hence the random variables $\int_0^t g(s)\, dY(s;\omega)$ are measurable with respect to the smallest sigma algebra generated by $\{Z_0(s;\omega), s \leqq t\}$, and hence $B(t)$ is contained in that algebra. This proves (2.25) since obviously the right side of (2.25) is contained in $B(t)$.

THEOREM 2.8.

$$Z_s(t;\omega) = \int_0^t (P(s)C^*(s) + B(s)D(s)^*)(D(s)D(s)^*)^{-1}\, dZ_0(s;\omega). \tag{2.28}$$

Proof. First of all, using (2.5), taking $Z_1(t;\omega)$ therein to be $Z_s(t;\omega)$, and $Z_2(t;\omega)$ to be $Z_0(t;\omega)$, and making use of (2.12), (2.15), we have that the conditional expectation of $Z_s(t;\omega)$ with respect to the smallest sigma algebra generated by $\{Z_0(s;\omega), s \leqq t\}$ is given by the right side of (2.28). But this algebra, by Theorem 2.7, is the same as $B(t)$, and $Z_s(t;\omega)$ is of course measurable with respect to $B(t)$. Hence (2.28) follows. Finally we note the following corollary.

COROLLARY 1. *Let* $\hat{R}(t) = E(\hat{x}(t;\omega)\hat{x}(t;\omega)^*)$. *Then* $\hat{R}(t)$ *is absolutely continuous, and*

$$\begin{aligned}\dot{\hat{R}}(t) = {} & A(t)\hat{R}(t) + \hat{R}(t)A^*(t) \\ & + (P(t)C(t)^* + B(t)D(t)^*)(D(t)D(t)^*)^{-1}(C(t)P(t) + D(t)B(t)^*).\end{aligned} \tag{2.29}$$

Proof. We have only to note that by the theorem,

$$\hat{x}(t;\omega) = \int_0^t A(s)\hat{x}(s;\omega) + \int_0^t (P(s)C(s)^* + B(s)D(s)^*)(D(s)D(s)^*)^{-1}\, dZ_0(s;\omega) \tag{2.30}$$

and the result follows by use of (2.12).

COROLLARY 2. $P(t)$ *is absolutely continuous,* $P(0) = 0$ *and*

$$\begin{aligned}\dot{P}(t) = {} & A(t)P(t) + P(t)A(t)^* + B(t)B(t)^* - (P(t)C(t)^* + B(t)D(t)^*)(D(t)D(t)^*)^{-1} \\ & \cdot(C(t)P(t) + D(t)B(t)^*).\end{aligned} \tag{2.31}$$

Proof. We have only to note that

$$P(t) = E(x(t;\omega)x(t;\omega)^*) - E(\hat{x}(t;\omega)\hat{x}(t;\omega)^*)$$

and by use of (2.29), the result follows.

The equations (2.27), (2.30), (2.31) are the Kalman filtering equations. From (2.30) it follows that $\hat{x}(t;\omega)$ can be determined to be continuous in t with probability one, and hence, so can $e(t;\omega)$, in particular.

REFERENCES

[1] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, Academic Press, New York, 1970.

[2] R. S. BUCY AND P. JOSEPH, *Filtering for Stochastic Process with Applications to Guidance*, Interscience, New York, 1968.

[3] T. KAILATH, *An innovations approach to least squares estimation*, IEEE Trans. Automatic Control, 13 (1968), pp. 646–655.
[4] E. NELSON, *Dynamical Theories of Brownian Motion*, Princeton University Press, Princeton, 1967.
[5] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.